Particle and microorganism enumeration data: Enabling quantitative rigor and judicious interpretation

Monica B. Emelko[*,†], *Philip J. Schmidt*[†], *Park M. Reilly*[‡]

Department of Civil and Environmental Engineering and Department of Chemical Engineering,

University of Waterloo, Waterloo, Ontario N2L 3G1

SUPPORTING INFORMATION

Number of pages = 7

Number of tables = 2

Number of figures = 1

Contents:

[*] Corresponding author phone: (519) 888-4567 x. 32208; fax: (519) 888-4349;
e-mail: mbemelko@civmail.uwaterloo.ca

[†] Civil & Environmental Engineering

[‡] Chemical Engineering

Derivation of the Negative Binomial Model

If the number of observed particles (x) in a sample of volume (V) taken from a source with homogeneous concentration (c) and enumerated by a method with recovery constant (p) is Poisson-distributed with mean (cVp), and the recovery constant is gamma-distributed with parameters (α,β), then the joint distribution $f(x,p)$ can be written as eq S1.

$$f(x,p) = \left[\frac{e^{-cVp}(cVp)^x}{x!}\right]\left[\frac{1}{\beta^\alpha \Gamma(\alpha)} p^{\alpha-1} e^{-p/\beta}\right] \quad \text{(S1)}$$

The marginal distribution for the number of observed particles (eq S2) can be derived as follows.

$$f(x) = \int_0^\infty \left[\frac{e^{-cVp}(cVp)^x}{x!}\right]\left[\frac{1}{\beta^\alpha \Gamma(\alpha)} p^{\alpha-1} e^{-p/\beta}\right] dp$$

$$f(x) = \frac{(cV)^x}{x!\beta^\alpha \Gamma(\alpha)} \int_0^\infty p^{x+\alpha-1} e^{-p\left(cV+\frac{1}{\beta}\right)} dp$$

$$f(x) = \frac{(cV)^x \Gamma(x+\alpha)}{x!\beta^\alpha \Gamma(\alpha)} \left(\frac{\beta}{cV\beta+1}\right)^{x+\alpha}$$

$$f(x) = \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \left(\frac{cV\beta}{cV\beta+1}\right)^x \left(\frac{1}{cV\beta+1}\right)^\alpha \quad \text{(S2)}$$

Summary of Model Components

Table S1. Summary of model components and sources of error

	Beta-Poisson Model	**Negative Binomial Model**
Representative Sampling Error	N/A	N/A
Random Sampling Error*	Poisson	Poisson
Analytical Error	Binomial (implicit)	Included in Poisson
Nonconstant Analytical Recovery	Beta	Gamma
Counting Errors (except false-positive)	Included in beta if recovery <100%	Included in gamma

* Alternative models using negative binomial random sampling error are not addressed.

Analysis of Data with Different Recovery Parameters

Equations 3-8, 10, and 13 assume that all samples regarded as replicates have the same recovery parameters (a,b or α,β). Modification of these equations to account for repeated samples that have different recovery parameters (e.g., due to variations in methodology or sample-specific recovery estimates) is very simple. In each of these equations, sample-specific recovery parameters can be provided (e.g., a_i,b_i).

Petterson, et al. (*S1*) considered sample-specific recovery estimates, in which precisely known numbers of pre-stained oocysts were seeded into environmental samples to concurrently evaluate recovery and indigenous oocyst concentrations. Such detailed recovery information can easily be incorporated into the beta-Poisson model presented herein using the recovery parameters (a_i,b_i) = (x_i^* + 1, $n_i^* - x_i^*$ +1) where n^* is the number of seeded particles and x^* is the number of seeded particles that were observed. These parameter values are obtained using a binomial model and Bayes' theorem (with uniform improper prior) as shown in Equation S3.

$$f(p_i) \propto \frac{n_i^*!}{x_i^*!(n_i^* - x_i^*)!} p_i^{x_i^*}(1-p_i)^{n_i^* - x_i^*} \rightarrow p_i : BETA(x_i^* + 1, n_i^* - x_i^* + 1) \quad \text{(S3)}$$

Calculating Posterior Distributions Using Numerical Integration

Given the likelihood functions represented by eqs 3 and 4 and an improper uniform prior, the marginal posterior probability density functions for concentration for the beta-Poisson and negative binomial models can be written explicitly as eqs S4 and S5, respectively.

Beta-Poisson:
$$f(c) = \frac{\prod_{i=1}^{r} c^{x_i} \cdot \int_0^1 e^{-cV_i p_i} p^{x_i + a - 1}(1-p)^{b-1} dp}{\int_0^\infty \left(\prod_{i=1}^{r} c^{x_i} \cdot \int_0^1 e^{-cV_i p_i} p^{x_i + a - 1}(1-p)^{b-1} dp \right) dc} \quad \text{(S4)}$$

Negative binomial:
$$f(c) = \frac{\prod_{i=1}^{r} \frac{c^{x_i}}{(cV_i\beta + 1)^{x_i + \alpha}}}{\int_0^\infty \left(\prod_{i=1}^{r} \frac{c^{x_i}}{(cV_i\beta + 1)^{x_i + \alpha}} \right) dc} \quad \text{(S5)}$$

These integrals cannot be solved explicitly, but can be approximated numerically. For example, the following integral can often be accurately approximated using 1000 intervals.

$$\int_0^1 e^{-cV_i p_i} p_i^{\,x_i+a-1}(1-p_i)^{b-1}\,dp_i$$

$$\approx \sum_{k=1}^{1000} e^{-cV_i\left(\frac{k-0.5}{1000}\right)}\left(\frac{k-0.5}{1000}\right)^{x_i+a-1}\left(\frac{1000.5-k}{1000}\right)^{b-1}(0.001)$$

$$= \sum_{k=1}^{1000}\exp\left((x_i+a+b-1)\ln(0.001) - cV_i\left(\frac{k-0.5}{1000}\right) + (x_i+a-1)\ln(k-0.5) + (b-1)\ln(1000.5-k)\right)$$

To approximate the denominator in either eq S4 or eq S5, a similar discrete approximation of the integral can be employed. This numerical approximation is more complicated because the upper boundary is infinite and because an appropriate step-size must be used in the discretization of the integral. The upper boundary should be chosen so that truncation error is minimal, and two step-sizes should be used to confirm that the numerical integration is converging on the correct value (a narrower step-size is needed if the result depends on the step-size). Numerical integration has been observed to be efficient and robust except when near-zero recovery values are common: then, the upper tail of the likelihood function narrows very slowly because a method with near-zero recovery yields very little information about concentration.

Gibbs Sampling Code for Visual Basic

The following code provides a framework for Gibbs sampling (from the beta-Poisson model) using the Visual Basic Editor in Microsoft Excel. The user must (1) modify the code to input data, (2) provide functions for random number generation, and (3) program data output and analysis (e.g., calculation of summary statistics such as the mean, mode, standard deviation, and 95% credible interval).

```
'    Declare input variables
     Dim inR as Integer          '    Number of replicate enumeration data
     Dim inX() as Integer        '    Array of sample counts (1 to inR)
     Dim sgV() as Single         '    Array of sample volumes (1 to inR)
     Dim sgA as Single           '    Beta distribution recovery parameter
     Dim sgB as Single           '    Beta distribution recovery parameter
     Dim lnNumBurn as Long       '    Number of burn-in iterations
     Dim lnNumSave as Long       '    Number of iterations to save in posterior

'    Declare temporary/output variables
     Dim inI as integer          '    Sample index
     Dim inN() as Integer        '    Array of unknown true counts (1 to inR)
     Dim sgP() as Single         '    Array of unknown recoveries (1 to inR)
     Dim lnSumN as Long          '    Sum of unknown true counts
     Dim sgSumV as Single        '    Sum of sample volumes
     Dim sgC() as Single         '    Markov chain of posterior conc. values
     Dim lnTrial as Long         '    Gibbs sampling iteration index
     Dim sgCV as Single          '    Product of concentration and Sum(Volume)
     Dim sgConc As Single        '    Temporary concentration value
     Dim sgLamda as Single       '    Temporary Poisson parameter

'    Input data (code not shown)

'    Select initial values for parameters
     ReDim inN(1 to inR)
     ReDim sgP(1 to inR)
     lnSumN = 0
     sgSumV = 0
     For inI = 1 to inR
          inN(inI) = Round(inX(inI) * (sgA + sgB) / sgA)
          lnSumN = lnSumN + inN(inI)
          sgSumV = sgSumV + sgV(inI)
     Next

'    Run Gibbs sampling
     ReDim sgC(1 to lnNumSave)
     For lnTrial = 1 to lnNumBurn + lnNumSave
          sgCV = GAMMA(lnSumN + 1)
          sgConc = sgCV / sgSumV
          If lnTrial > lnNumBurn Then
               sgC(lnTrial – lnNumBurn) = sgConc
          End If
          lnSumN = 0
          For inI = 1 to inR
               sgP(inI) = BETA(inX(inI) + sgA, inN(inI) – inX(inI) +sgB)
               sgLamda = sgConc * sgV(ini) * (1 – sgP(inI))
               inN(inI) = POISSON(sgLamda) + inX(inI)
               lnSumN = lnSumN + inN(inI)
          Next
     Next
```

Convergence and Mixing

Convergence and mixing are important considerations when using Markov chains (*S2*). Gibbs sampling has converged upon the posterior distribution when it produces a sequence of values that are collectively representative of the posterior distribution and not influenced by the initial values of the Markov chain. An appropriate number of iterations must be discarded at the outset of Gibbs sampling ("burn-in") to ensure that the first recorded value is essentially a random sample from the posterior distribution (i.e., it is unaffected by the supplied initial values). Mixing corresponds to the degree of serial correlation in a Markov chain; poorly mixing chains (i.e., with high serial correlation) will converge slowly.

Figure S1a shows an example of Gibbs sampling output using the beta-Poisson model with the following input: x_1 = 376, x_2 = 388, V_1 = 10 L, V_2 = 10 L, a = 287.08, b = 94.76. Rather than using the initial values for $\{n_i\}$ that are proposed in Figure 1a (which ensure rapid convergence), each was set to 750. Despite the highly improbable initial values, the Markov chain has converged in fewer than 50 iterations. Figure S1a illustrates excellent mixing.

Poorer mixing has been observed when near-zero recovery is common. Figure S1b shows an example of Gibbs sampling output using the beta-Poisson model with the following input: x_1 = 3, V_1 = 100 L, a = 4, b = 6. Once again, n = 750 was input to evaluate convergence. This figure shows slower convergence (a burn-in of approximately 500 iterations appears to be appropriate) and poor mixing.

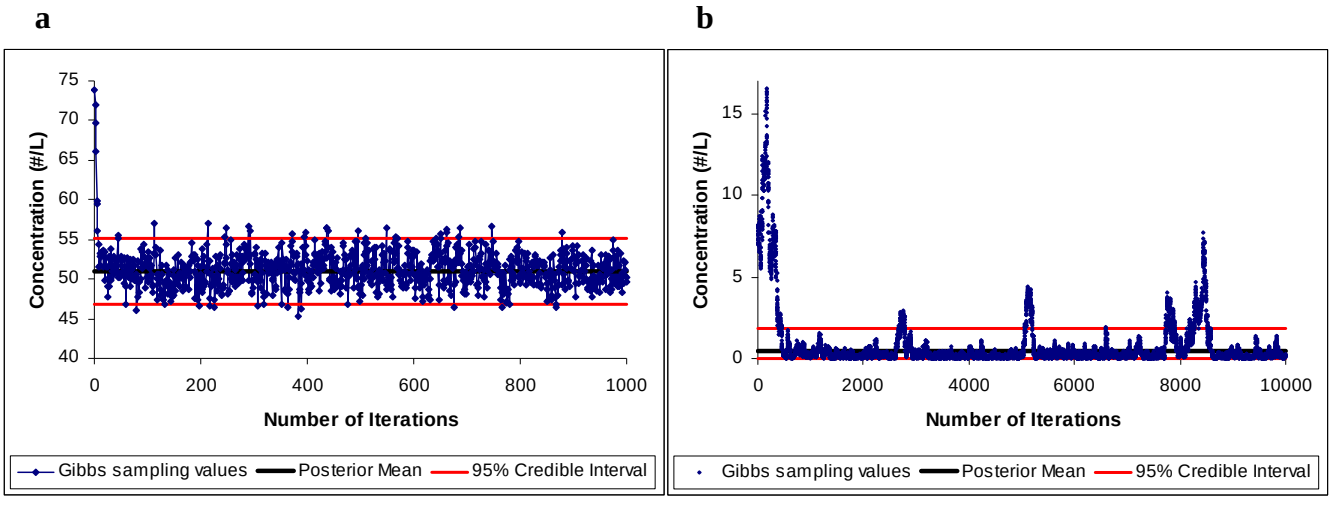

Figure S1. Convergence and mixing of Gibbs sampling algorithm

Sample Enumeration Data Used in Figure 2

Table S2. Sample enumeration data

	Initial		Final			
Volume (L)	10	10	50	50	50	50
Count	376	388	16	16	19	29

c_I = 50 microorganisms/L, c_F = 0.5 microorganisms/L, (a,b) = (287.08, 94.76)

Literature Cited

(*S1*) Petterson, S. R.; Signor, R. S.; Ashbolt, N. J. Incorporating method recovery uncertainties in stochastic estimates of raw water protozoan concentrations for QMRA. *J. Water Health* **2007**, 5, 51-65.

(*S2*) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian Data Analysis*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, 2004.