OPEN ACCESS Freely available online

PLOS ONE

Bayesian Inference from Count Data Using Discrete Uniform Priors

Federico Comoglio[1☯], **Letizia Fracchia**[2], **Maurizio Rinaldi**[2*☯]

1 Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology Zürich, Basel, Switzerland, **2** Dipartimento di Scienze del Farmaco, Università degli Studi del Piemonte Orientale "Amedeo Avogadro", Novara, Italy

Abstract

We consider a set of sample counts obtained by sampling arbitrary fractions of a finite volume containing an homogeneously dispersed population of identical objects. We report a Bayesian derivation of the posterior probability distribution of the population size using a binomial likelihood and non-conjugate, discrete uniform priors under sampling with or without replacement. Our derivation yields a computationally feasible formula that can prove useful in a variety of statistical problems involving absolute quantification under uncertainty. We implemented our algorithm in the R package dupiR and compared it with a previously proposed Bayesian method based on a Gamma prior. As a showcase, we demonstrate that our inference framework can be used to estimate bacterial survival curves from measurements characterized by extremely low or zero counts and rather high sampling fractions. All in all, we provide a versatile, general purpose algorithm to infer population sizes from count data, which can find application in a broad spectrum of biological and physical problems.

Citation: Comoglio F, Fracchia L, Rinaldi M (2013) Bayesian Inference from Count Data Using Discrete Uniform Priors. PLoS ONE 8(10): e74388. doi:10.1371/journal.pone.0074388

Editor: Zaid Abdo, Institution and Department: Agricultural Research Service, United States of America

Received June 17, 2013; **Accepted** July 31, 2013; **Published** October 7, 2013

Funding: This work was funded by the Ministry of Education, University and Research (MIUR, http://www.miur.it). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: maurizio.rinaldi@unipmn.it

☯ These authors contributed equally to this work.

Introduction

Absolute quantification of objects, namely the determination of their total number from measurements subject to sampling uncertainty, is a classical problem in statistical inference. In this work, we consider a finite population of identical objects homogeneously dispersed in a finite volume. We assume that measurable fractions of the volume can be sampled and that the number of objects therein can be counted. Given the resulting set of measurements, we address the problem of estimating the population size and its uncertainty using a Bayesian approach with least informative prior distribution.

In a Bayesian treatment of this problem, counts are usually considered to be either Poisson, binomial or negative binomial distributed, depending on the nature of the problem at hand. For example in genomics, over-dispersed sequence count data as those obtained by RNA-Seq are more effectively modeled by a negative binomial than by a Poisson distribution, as the former provides a more flexible mean-variance relationship [1–3]. When counts are modeled as a binomial distribution, the binomial likelihood is generally coupled to a conjugate prior to yield a closed form posterior distribution, which corresponds to a simple update of the prior parameters. However, handy computations do not imply that the prior distribution correctly encodes our prior belief, which instead requires specification of both the class of prior distributions and parameters. This choice is paramount when dealing with limited sample sizes [4–6], which typically affect biologically relevant inference processes. In addition, in many applications we often have no ground to expect certain simple events to be more likely to occur than others. Therefore, as there is no reason to prefer one distribution over another, a uniform prior distribution can be used to encode this prior belief. This is a formulation of the so called principle of indifference [7], also known as Laplace's principle of insufficient reason [8]. Here, we resort to this principle in order to propose a Bayesian approach in which we introduce the least prior information over a discrete sample space. As the principle of indifference considers each possible outcome as equiprobable, it naturally leads to discrete uniform priors, a class of maximum entropy priors on a discrete sample space [9–11]. However, in order to make use of this class of prior distributions for Bayesian inference, we had to address two specific issues: i) a discrete uniform prior with infinite support is an improper prior and ii) it is not a conjugate prior for neither of the above mentioned likelihoods for counts data. Although improper priors are argument of long-standing debate in the field, Jaynes [10] provided a rigorous advice on how to use improper prior for Bayesian inference. Therefore, we addressed the first issue by following Jaynes's approach [10], namely we considered a well defined limit of discrete uniform priors and verified that, even in the limit, the resulting posterior is a proper probability distribution.

Next, despite non-conjugacy we were able to obtain a computationally tractable formula for the posterior distribution of the population size using a binomial likelihood. Particularly, we analyzed two different sampling schemes where objects are

either drawn with or without replacement and report a formula for the posterior distribution for each of these cases.

We implemented our algorithms in the R package dupiR and as a showcase, we applied our framework to microbial count data obtained through viable plate counts. A number of studies in clinical and environmental microbiology, and food safety, deal with the quantitative determination of bacteria. Interestingly, low bacterial loads in a sample can challenge bacteria enumeration methods because irrespective of the sampling fraction, they result in low viable counts that are generally considered to be statistically unreliable and hence discarded. By analyzing bacteria survival data exhibiting extremely low counts and rather high sampling fractions, we show that our approach is able to cope well with these data, providing reliable credible intervals for the total number of bacteria even in such extreme cases.

Results

General concepts and notation

We consider a finite volume V containing n identical and uniformly distributed objects. A single count of k objects from a sampling fraction r, with $0 \le r \le 1$, is initially considered (Figure 1A). Our goal is to estimate n using a class of discrete uniform priors. Here, counts follow a binomial distribution $\mathbf{B}[n,r]$

$$P(k|n,r) = \mathbf{B}[n,r](k) = \binom{n}{k} r^k (1-r)^{n-k}$$

and by Bayes' rule

$$P(n|k,r) = P(k|n,r)\frac{P(n|r)}{P(k|r)}.$$

We assume that our prior belief on n does not depend on r, namely $P(n|r) = P(n)$, and that $P(n)$ is the discrete uniform distribution with support $\{n_1, n_1+1, \ldots, n_2\}$ given by

$$P(n) = \mathbf{U}[n_1,n_2](n) = \frac{1}{n_2 - n_1 + 1}, \quad n_1 \le n \le n_2. \tag{1}$$

In the following, we consider the general case in which we are given m measurements $k_1, \ldots, k_m$ from sampling fractions $r_1, \ldots, r_m$ (Figure 1B) and derive a formula for the posterior distribution $P(n|k_1, \ldots, k_m, r_1, \ldots, r_m)$ distinguishing between two sampling schemes: i) sampling with replacement; ii) sampling without replacement.

Derivation of the posterior distribution under sampling with replacement

Here, we derive $P(n|k_1, \ldots, k_m, r_1, \ldots, r_m)$ given sample counts drawn with replacement. Assuming n to be conditionally independent of $r_1, \ldots, r_m$, from Bayes' rule we have

$$P(n|k_1, \ldots, k_m, r_1, \ldots, r_m) = P(k_1, \ldots, k_m|n, r_1, \ldots, r_m)\frac{P(n)}{P(k_1, \ldots, k_m|r_1, \ldots, r_m)}. \tag{2}$$

Assuming that the measurements are independent of each other and that counts are conditionally independent of the sample fractions the likelihood factorizes to $P(k_1, \ldots, k_m|n, r_1, \ldots, r_m) = \Pi_{i=1}^m P(k_i|n, r_i)$ and therefore equation 2 can be written as:

$$P(n|k_1, \ldots, k_m, r_1, \ldots, r_m) = \frac{\Pi_{i=1}^m P(k_i|n, r_i) P(n)}{\sum_n \Pi_{i=1}^m P(k_i|n, r_i) P(n)}.$$

Let $P(n) = \mathbf{U}[n_1,n_2](n)$ as introduced in equation 1. Then

$$P(n|k_1, \ldots, k_m, r_1, \ldots, r_m) = \frac{\Pi_{i=1}^m P(k_i|n, r_i) U[n_1,n_2](n)}{\sum_{n=n_1}^{n_2} \Pi_{i=1}^m P(k_i|n, r_i) U[n_1,n_2](n)}. \tag{3}$$

As the interval $[n_1,n_2]$ can be arbitrarily large, the denominator of equation 3:

$$P(k_1, \ldots, k_m|r_1, \ldots, r_m) = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} \prod_{i=1}^m \binom{n}{k_i} r_i^{k_i} (1-r_i)^{n-k_i}$$

features a potentially intractable summation over the prior support. To address this issue we introduce the following lemma.

Lemma 1. Let $k = (k_1, \ldots, k_m)$, and $x = \Pi_{i=1}^m (1-r_i)$. For $n_2 \ge \max(k)$

$$\sum_{n=n_1}^{n_2} \Pi_{i=1}^m \binom{n}{k_i} r_i^{k_i} (1-r_i)^{n-k_i} = \Pi_{i=1}^m \left(\frac{r_i}{1-r_i}\right)^{k_i} (F(k,n_1,x) - F(k,n_2+1,x)) \tag{4}$$

where

$$F(k,n,x) = \sum_{t_1=0}^{k_1} \cdots \sum_{t_m=0}^{k_m} \prod_{i=1}^m \binom{n+T_{i-1}}{k_i - t_i}_i \binom{T}{t_i} \frac{x^{n+T}}{(1-x)^{1+T}} \tag{5}$$

and $T_i = \sum_{j=1}^i t_j$, $T = T_m$, and $T_0 = 0$. The proof is provided in the Appendix (see Text S1). Based on the fact that the sample counts are generally orders of magnitude smaller than the population size, this lemma allows to replace the sum over n by nested sums over k_i, with $i \in \{1,2,\ldots,m\}$. Although the computational complexity of equation 4 is $O(\max(k)^m)$, the number of measurements is typically limited in a number of practical applications, thus enabling direct computation of the expression.

Using Lemma 1 we can express the posterior distribution $P(n|k_1, \ldots, k_m, r_1, \ldots, r_m)$ as follows.

Theorem 1. If $P(n) = U[n_1,n_2](n)$ then,

$$P(n|k_1, \ldots, k_m, r_1, \ldots, r_m) = \frac{x^n \Pi_{i=1}^m \binom{n}{k_i}}{F(k,n_1,x) - F(k,n_2+1,x)} \tag{6}$$

where k,x are defined as in Lemma 1.

Proof. The proof follows by rewriting the posterior distribution of n (equation 3) as

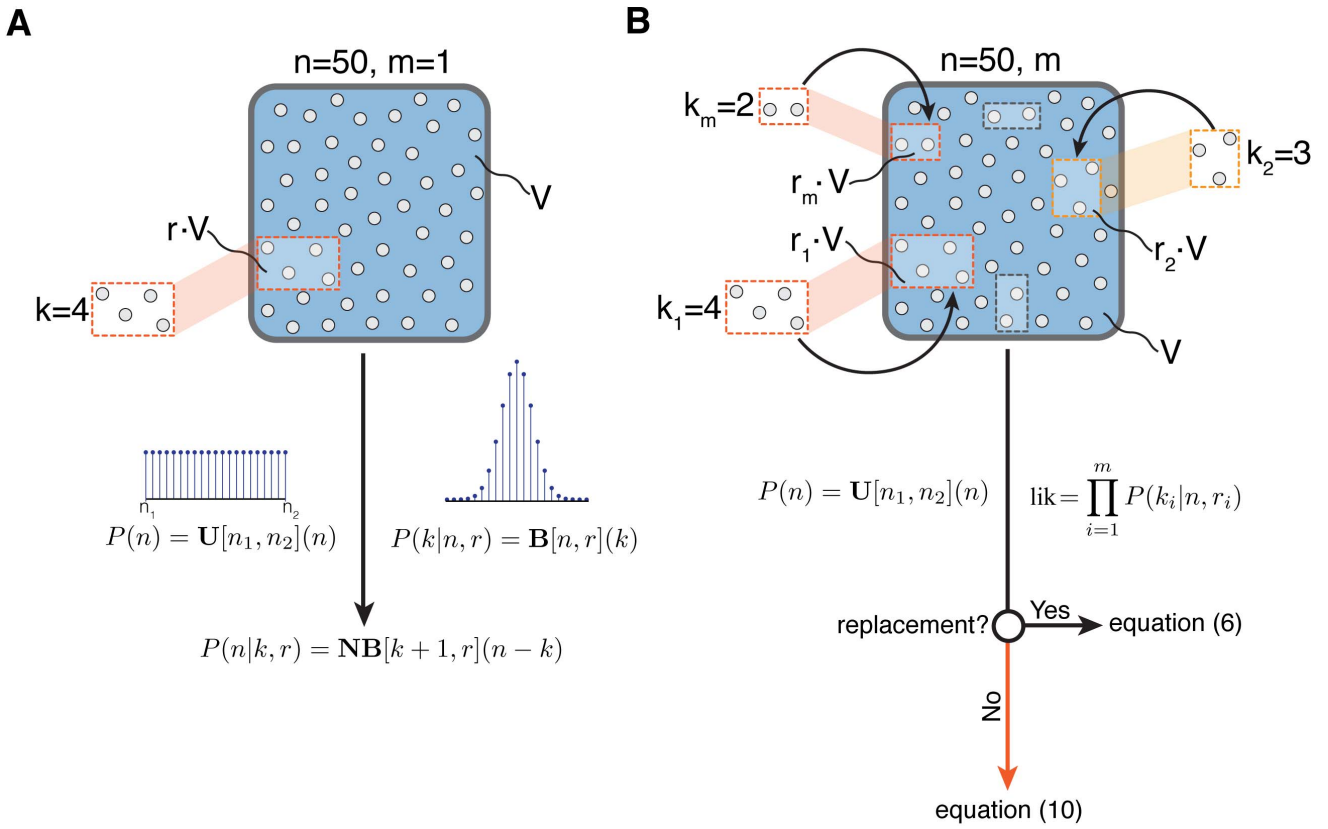

Figure 1. Schematic representation of the problem and of our inference framework. (A) A total of n identical objects (in gray, $n = 50$ in this example) is homogeneously dispersed in a finite volume V. A fraction r of V, having volume rV, is sampled (dashed red rectangle) and the number of object therein, denoted with k ($k = 4$ in this example) is determined. Given the measurement, the posterior distribution of n is a negative binomial probability distribution $P(n|k,r)$ (bottom) computed from a binomial likelihood $P(k|n,r)$ (right) and a discrete uniform prior $P(n)$ (left). (B) Generalization of (A) to m measurements. Fractions of volume $r_i V$ are sampled the number of objects therein (k_i) determined as before. However, when $m>1$ two cases can be distinguished: i) the fractions are replaced (sampling with replacement); ii) the fractions are removed from V (sampling without replacement). In both cases, we derived a formula for the posterior distribution which is reported in the text as equation 6 for case i and equation 10 for case ii.
doi:10.1371/journal.pone.0074388.g001

$$P(n|k_1,\ldots,k_m,r_1,\ldots,r_m) = \frac{\prod_{i=1}^{m}\binom{n}{k_i} r_i^{k_i}(1-r_i)^{n-k_i}}{\prod_{i=1}^{m}(r_i/(1-r_i))^{k_i}\left(F(k,n_1,x)-F(k,n_2+1,x)\right)} = \frac{x^n \prod_{i=1}^{m}\binom{n}{k_i}}{F(k,n_1,x)-F(k,n_2+1,x)}$$

Corollary 1. Suppose $n_1 = 0$. Then $P(n) = \mathbf{U}[0,n_2](n) = \frac{1}{n_2+1}$, for $0 \le n \le n_2$, and we have

$$P(n|k_1,\ldots,k_m,r_1,\ldots,r_m) = \frac{x^n \prod_{i=1}^{m}\binom{n}{k_i}}{F(k,\max(k),x)-F(k,n_2+1,x)}$$

$0 \le n \le n_2$.

Corollary 2. If $P(n) = \mathbf{U}[0,n_2](n)$ for $0 \le n \le n_2$ then in the limit $n_2 \to \infty$

$$P(n|k_1,\ldots,k_m,r_1,\ldots,r_m) = \frac{x^n \prod_{i=1}^{m}\binom{n}{k_i}}{F(k,\max(k),x)}.$$

Notice that if a single measurement is given ($m = 1$), the posterior probability of n reduces to

$$P(n|k,r) = \frac{\binom{n}{k} r^{k+1}(1-r)^{n-k}}{1-\sum_{t=0}^{k}\binom{n_2+1}{t} r^t (1-r)^{n_2+1-t}}. \quad (7)$$

(see Appendix in Text S1). Let $n = j+k$. In the limit $n_2 \to \infty$ we obtain

$$P(n|k,r) = \binom{n}{k} r^{k+1}(1-r)^{n-k} = \binom{j+k}{j} r^{k+1}(1-r)^j \quad (8)$$
$$= \mathbf{NB}[k+1,r](j) = \mathbf{NB}[k+1,r](n-k)$$

namely, the posterior $P(n|k,r)$ is a negative binomial distribution shifted by k units and parametrized by $k+1$ and r.

Derivation of the posterior distribution under sampling without replacement

Suppose that m fractions of the volume V are sampled uniformly at random without replacement. Let $k_1,\ldots,k_m$ be ordered sample counts, drawn from sampling fractions $r_1,\ldots,r_m$ computed with respect to V. Clearly, if $m=1$ the posterior $P(n|k_1,r_1)$ is given by equation 7 and by working in the limit $n_2\rightarrow\infty$ (equation 8) we have

$$P(n|k_1,r_1)=\mathbf{NB}[k_1+1,r_1](n-k_1).$$

Consider now a second measurement sampled from a fraction r_2 of V and therefore equal to a fraction $\bar{r}_2=\frac{r_2}{1-r_1}$ of the residual volume $\bar{V}$. In this case, the likelihood is given by $\mathbf{B}[n-k_1,\bar{r}_2]$ and the prior is $\mathbf{NB}[k_1+1,1-r_1](n-k_1)$. Therefore, the posterior distribution of n is given by

$$\begin{aligned}&P(n|k_1,k_2,r_1,r_2)=\mathrm{NB}[k_2+k_1+1,1-(1-r_1)(1-\bar{r}_2)]\\&(n-k_1-k_2)=\mathrm{NB}[k_2+k_1+1,r_1+r_2)](n-k_1-k_2).\end{aligned}\tag{9}$$

Let $K=\sum_{i=1}^{m}k_i$ and $R=\sum_{i=1}^{m}r_i$. By induction, equation 9 can be generalized to m measurements, obtaining

$$\begin{aligned}&P(n|k_1,\ldots,k_m,r_1,\ldots,r_m)=\\&\mathrm{NB}[K+1,1-\Pi_{i=1}^{m}(1-\bar{r}_i)](n-K)\\&=\mathrm{NB}[K+1,R](n-K).\end{aligned}\tag{10}$$

This result has two important properties. First, the computation of the posterior distribution depends only on the sum of the counts and on the sum of the sampling fractions, becoming therefore independent on the number of measurements. As a consequence, any permutation of counts and fractions leads to the same posterior distribution. Second, there exists an equivalence between experiments yielding the same values of K and R through a different number of measurements, i.e. counting k_i in fractions r_i for $m>1$ yields the same posterior distribution as counting $\sum_{i=1}^{m}k_i$ counts in a fraction $\sum_{i=1}^{m}r_i$ in single measurement.

The R package dupiR

We implemented our algorithms as a package for the statistical environment R [12]. The package, which we called dupiR (discrete uniform prior-based inference with R) is available from the Comprehensive R Achive Network (CRAN) along with the relevant package manual. dupiR is based on the custom S4 class Counts, which is used to store sample information, statistical attributes and inference results. By default, the package assumes that samples have been drawn without replacement. Given a set of sample counts $\{k_1,\ldots,k_m\}$ and fractions $\{r_1,\ldots,r_m\}$, dupiR defines the default support interval for the discrete uniform prior distribution as $[0.5\cdot\hat{\boldsymbol{n}},2\cdot\hat{\boldsymbol{n}}]$, where $\hat{\boldsymbol{n}}$ is the maximum likelihood estimate of n computed as K/R, where $K=\sum_{i=1}^{m}k_i$ and $R=\sum_{i=1}^{m}r_i$. For the special case $K=0$, the prior support is defined as $[0,1/\min(\{r_1,\ldots,r_m\})]$. This setup proved to be effective across a variety of simulated measurements. However, the user can override default values by explicitly using the variables n1 and n2 to define a custom prior support.

Posterior distributions can be computed using the function computePosterior, where the logical parameter replacement specifies whether counts were sampled with or without replacement. Posterior parameters can be obtained using getPosteriorParam, which returns a point estimate of n equal to its maximum a posteriori (MAP) and the corresponding credible interval at a specified confidence level (default to 95%), among other parameters. Finally, dupiR can be used to produce publication-level quality figures representing posterior distributions and parameters simply via the plot function. Further information are provided in the package documentation.

Applications to bacterial enumeration

Absolute quantification of bacteria in biological samples is performed routinely for a broad spectrum of applications ranging from diagnostics to food analysis. A standard method for bacterial enumeration is the plate count method, which despite well-recognized limitations provides an indirect measure of cell density solely based on viable bacteria [13]. Viable plate counts - the discrete outcome of this method - are then generally used to compute point estimates of the bacterial concentration in the original sample. Although Bayesian estimates of the uncertainty associated to bacteria quantification have been previously proposed, these methods assume Poisson distributed microbial counts [14,15]. Particularly, Clough *et al.* [14] adopted a Poisson likelihood $\mathbf{Pois}[\lambda](k)=\frac{e^{-\lambda}\lambda^k}{k!}$ with rate $\lambda=rn$ and a Gamma prior distribution $\boldsymbol{g}[\kappa,\rho](n)=\frac{1}{\Gamma(\kappa)}\rho^{\kappa}n^{\kappa-1}e^{-\rho n}$, where κ and ρ are the shape and the rate parameters, respectively. It then follows that the posterior distribution $P(n|k_1,\ldots,k_m,r_1,\ldots,r_m)$ is itself a Gamma distribution given by

$$P(n|k_1,\ldots,k_m,r_1,\ldots,r_m)=g[\kappa+K,\rho+R](n)\tag{11}$$

where $K=\sum_{i=1}^{m}k_i$ and $R=\sum_{i=1}^{m}r_i$. Hereinafter, we will refer to this setup as the GP (Gamma-Poisson) method. In applying the GP method to bacteria enumeration, the authors chose $\kappa=1$ and $\rho=10^{-6}$. Notice that the gamma distribution is appropriate to model continuous variables and therefore a continuous approximation to n is assumed in this model.

By analyzing equation 11 we can observe that when $R\ll1$ and $\binom{n}{k}\sim\frac{n^k}{k!}$, the expression converges to our posterior distribution under sampling without replacement (equation 10) as

$$P(n|K,R)=\mathbf{NB}[K+1,R](n-K)=\binom{n}{K}R^{K+1}(1-R)^{n-K}.$$

Indeed notice that for small values of R the expression above depends on n as $n^Ke^{-Rn}\sim\boldsymbol{g}[K+1,R](n)$ and that by setting $\rho=0$, $\boldsymbol{g}[K+1,R](n)$ is equal to equation 11.

Convergence implies that for a broad range of measurements our inference framework and the GP method provide comparable results. However, when the difference between sampling methods is not negligible, i.e. when sampling fractions are large, the results provided by the two methods become significantly different. To investigate this difference in greater details and to assess the performances of our inference framework, we simulated measurements from total sampling fractions spanning two orders of magnitude and we compared the posterior distributions inferred with our method to those obtained via the GP method using the

Jensen-Shannon divergence (JS-divergence), a symmetric version of the Kullback-Leibler divergence (see Methods). Our simulation results show that when R is so small that the effect of replacement is negligible, posterior distributions computed using our method or with the GP method correspond to the same probability distribution for any practical purpose (Figure S1). More precisely, the effect of replacement can be neglected when $R \leq 1/32$, a value at which the JS-divergence between posterior distributions computed from sampling with and without replacement drops below 10^{-4} (Figure S2). However, when $R > 1/32$, the two approaches differ substantially. For these values of total sampling fractions, posterior distributions computed using our algorithm exhibit a lower variance than those computed with the GP method (Figure 2), thus providing narrower credible intervals. It is noteworthy to observe that this result is not a mere consequence of an inappropriate parametrization of the Gamma prior. Indeed, simulations performed by varying ρ over several orders of magnitude (from $\rho = 10^{-5}$ to $\rho = 0.1$) showed that differences between posterior distributions remain significant irrespective of ρ (Figure S3, A–E). Rather, as expected, extreme rate parameters can lead to posterior distributions that are dominated by prior belief (see Figure S3, F for an example), emphasizing the importance of an appropriate prior parametrization.

Taken together, these results underscore the generality of our inference method, which is able to cope with measurements derived from any range of K and R, including extreme total sampling fractions and counts. This latter property is desirable for bacterial enumeration. In fact, only measurements with $30 \leq K \leq 300$ are routinely used to infer the population size [16] and those localizing outside this range are currently discarded. Clearly, if $K < 30$ and $R \ll 1$ it is often easy to obtain a measurement with K falling within the recommended range simply by considering those samples in the dilution series which are less diluted (i.e. obtained from a higher sampling fraction). However, when n is small, measurements obtained from high sampling fractions can still yield low counts. Studies investigating bacterial survival upon physical or chemical treatments or in different environmental conditions are often confronted with this limitation. Bacterial survival studies are generally based on time-course bacterial enumeration using different experimental techniques and aim to estimate bacterial survival curves that in turn are used to compare cell viability across conditions. In a recent environmental microbiology study, Fracchia *et al.* investigated the suitability of biosolids as inoculum vehicle for the plant-growth promoting rhizobacteria *Pseudomonas fluorescens* [17]. Here, we deal with a single time series which was generated as described in [17] (six time points where for each sample at least five technical

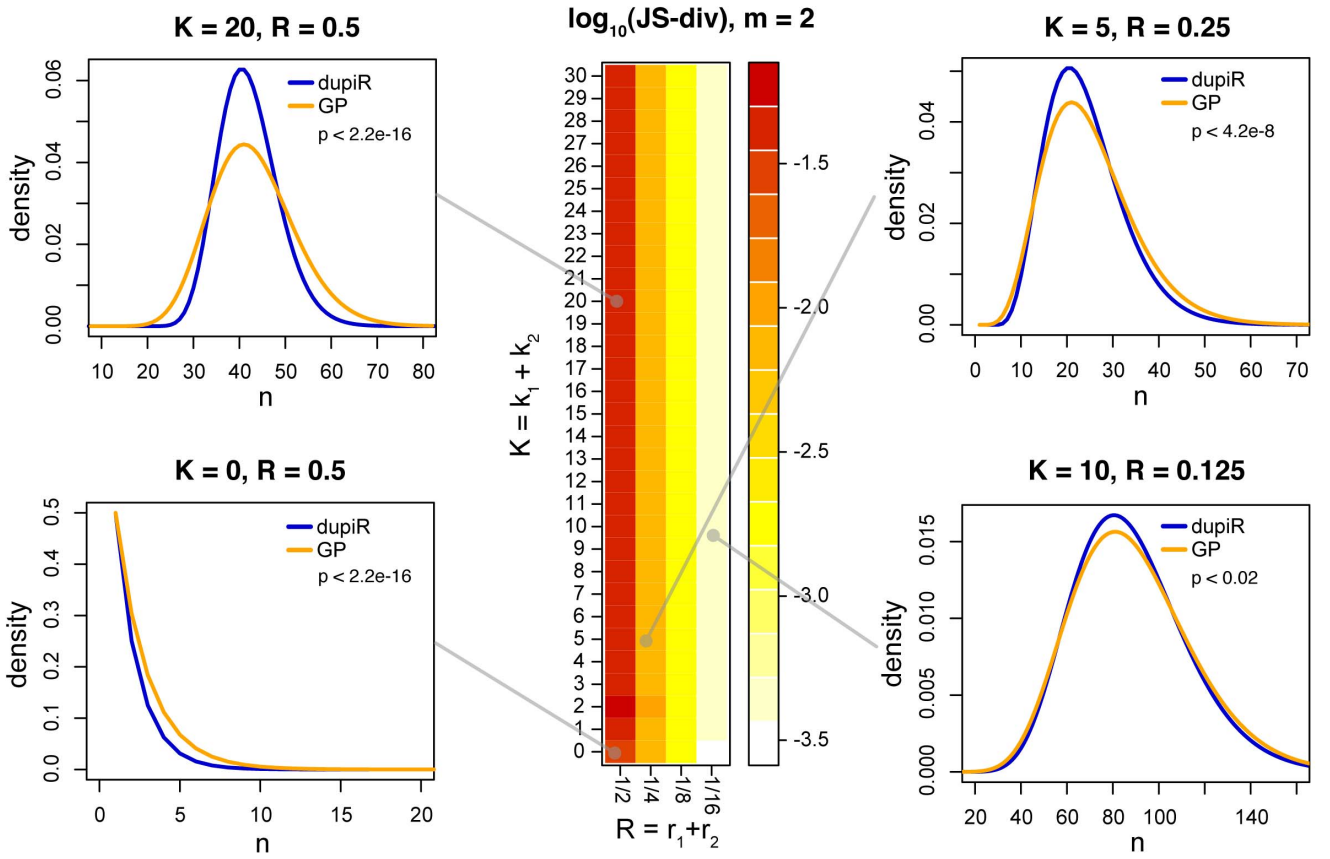

Figure 2. Comparison between posterior distributions computed with dupiR and with the GP method. Middle: JS-divergence (expressed in $\log_{10}$) between posterior distributions computed with our algorithm using sampling without replacement or with the GP method (Clough *et al.* [14]) as a function of total counts ($K \in \{0,1,\ldots,30\}$, see Methods) and total sampling fractions (R) obtained from two measurements ($m = 2$). Right and Left: examples of posterior distributions corresponding to values of (K,R) indicated by grey lines are illustrated. p-values have been computed using a two-sided Kolmogorov-Smirnov test.
doi:10.1371/journal.pone.0074388.g002

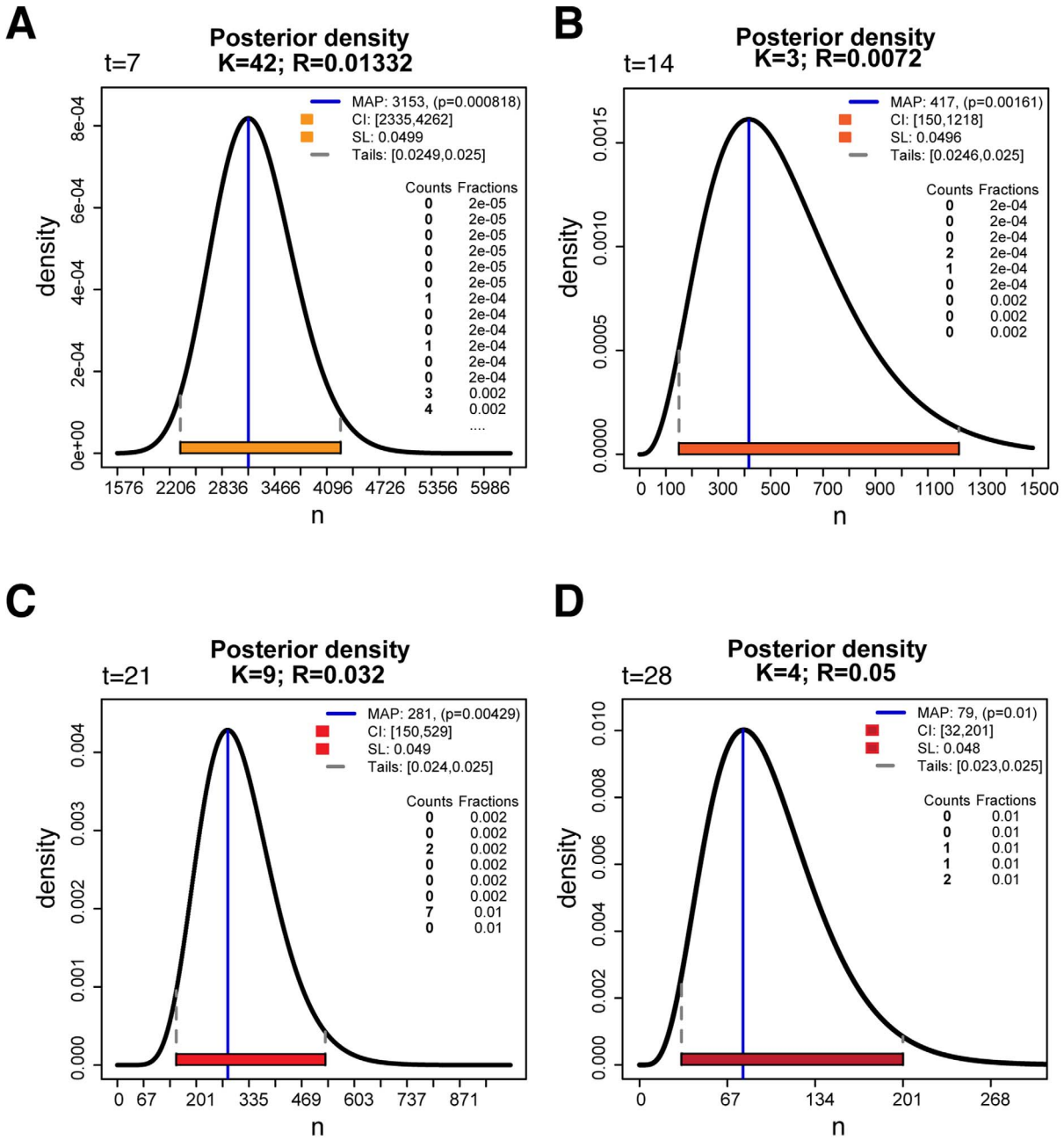

Figure 3. Examples of dupiR graphical output. Examples of posterior distributions of the population size *n* estimated and plotted with dupiR for time points (A) $t = 7$, (B) $t = 14$ (C) $t = 21$ and (D) $t = 28$ days. By default, the graph of the posterior distribution (solid black line) is plotted along with a statistical summary containing the maximum a posteriori (MAP, indicated by the blue vertical line) of *n*, the corresponding credible intervals (CI, green and dashed grey lines) at a significance level (SL) of 0.05 and the tails probability of the distribution function.
doi:10.1371/journal.pone.0074388.g003

replicates were subjected to bacterial enumeration, see Methods) where only the first two time points yielded $30 \leq K \leq 300$ and where more than 50% of the measurements in later time points showed $K = 0$ (see Figure 3). Instead of discarding these measurements, we applied dupiR to compute posterior distributions and estimated the maximum a posteriori (MAP) of *n* from all time points. These values were then used to fit a power-law model (see Methods) that shows good agreement with the experimental data (residual standard error of 0.1269 on 3 degrees of freedom), thus enabling us to estimate a survival curve of *P. fluorescens* in a time series characterized by extremely low viable counts (Figure 4). Clearly, dupiR estimates can be integrated into more complex models of cell growth or survival for which several mathematical approaches have been proposed [18–21].

Discussion

Parametrization of the prior probability distribution is a key step in Bayesian statistics. This step requires particular care for small sample sizes, as posterior distributions can be easily dominated by prior belief unless the parameters reflect an appropriate equivalent sample size of the prior distribution [4,5]. In addition, the choice

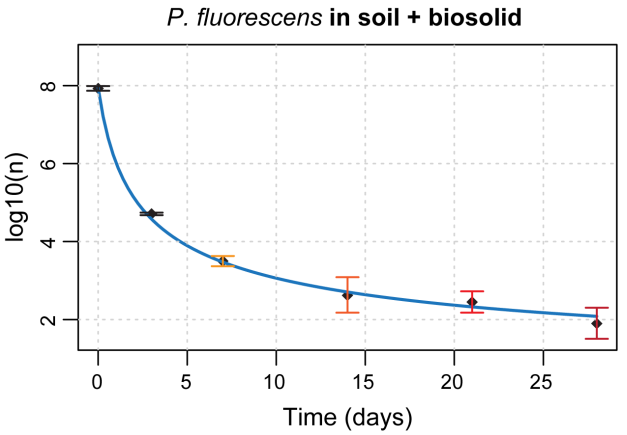

Figure 4. Application of discrete uniform priors to bacterial survival curves estimation. Estimated bacterial survival curve (light blue line, see Methods) of *P. fluorescens* inoculated in soil supplemented with biosolid. Time points from $t=7$ to $t=28$, characterized by zero or extremely low viable counts, are indicated in tones of red and the corresponding posterior distributions are shown in Figure 3.
doi:10.1371/journal.pone.0074388.g004

of the prior is sometimes driven by convenience rather than prior belief. In this study, we set out to overcome these intrinsic limitations by implementing Keynes' principle of indifference [7] in a Bayesian framework to infer population sizes (n) from sample measurements. Notably, we did not limit ourselves to a theoretical treatment of the subject but we provide an optimized, general purpose implementation of our algorithm in the R package dupiR.

By attributing equal probabilities to each possible outcome, Keynes' principle of indifference is naturally encoded in discrete uniform priors. Notice that the application of this principle in our univariate, discrete problem is free from possible unexpected behavior that are known to arise in multivariate, continuous applications (e.g. see [22]). Although discrete uniform priors are not conjugate for likelihoods commonly adopted in dealing with count data, we were able to derive the posterior probability distribution of n using a binomial likelihood. If data are obtained through sampling with replacement, we report a computationally tractable formula of the posterior distribution of n which could be obtained by converting a summation over the prior support to multiple summations over the range of sample counts only. Indeed, while the former can be theoretically unbound, sample counts are typically orders of magnitude smaller than n. The special case in which only a single measurement is available leads to a negative binomial posterior distribution, which was then used as a building block to extend our framework to an arbitrary number of measurements obtained from sampling without replacement. The properties of the posterior distributions we obtained depend on the sampling method. Particularly, while under sampling with replacement measurements contribute individually to the inference process, if no replacement is performed then the posterior distribution depends only on the total of sample counts and fractions. This property allows computations to be independent of the number of measurements.

The sampling method has no influence on the result if the total sampling fraction is modest compared to the total volume ($R \ll 1$). This holds true for typical experimental settings and under this condition the performances of our algorithm are comparable to those of other Bayesian methods reported in literature, such as the GP method [14] (Figures S1 and S2). However, the results of the two methods diverge when the effect of replacement can no longer be neglected. In this cases, our method provides posterior distributions that are characterized by a significantly smaller variance compared to those obtained using a Gamma prior (Figure 2). This property can be seen analytically. Since $\sigma^2_{\mathbf{NB}} = \frac{(K+1)(1-R)}{R^2}$ and $\sigma^2_g = \frac{K+1}{(\rho+R)^2}$, when $\rho \ll R$ we have $\sigma_{\mathbf{NB}}/\sigma_g \sim (1-R)$ and hence $\sigma_g > \sigma_{\mathbf{NB}}$.

We showed an application of our method in the context of bacterial enumeration, where we investigated the survival of an engineered strain of *P. fluorescens* using a time series with very low or zero viable counts and rather high sampling fractions. Although in this work we dealt with viable plate counts only, data generated by other laboratory techniques, such as the direct count [23] and the drop plate method [24] can be analyzed with dupiR. In addition, combining our algorithm with automatic plate counting [25] could result in a reliable and robust pipeline for bacteria enumeration via plate counting methods.

All in all, we provided a general purpose algorithm to infer population sizes from count data. We believe that the method can be applied to a broad spectrum of applications in both biological and physical sciences.

Materials and Methods

Simulation

Given a set of total counts K and a set of total sampling fractions R we considered the pairs $K \times R$ and computed posterior distributions using either our posterior formula under sampling without replacement or the GP method. For each $r \in R$, all posterior distributions were computed using the same discrete uniform prior by setting its support to the interval $[0, 2 \cdot \hat{n}_{\max}]$, where $\hat{n}_{\max} = \frac{\max(K)}{r}$. Posterior distributions computed via the two methods were compared by computing the Jensen-Shannon divergence (JS-divergence), a symmetrised Kullback-Leibler divergence (KL-divergence) [26]. Given two discrete probability distributions p and q, the KL-divergence (in bits) is defined as

$$\mathrm{KL}(p,q) = \sum_i p_i \log_2 \left(\frac{p_i}{q_i}\right).$$

Since the KL-divergence is not symmetric, different symmetrization procedures have been proposed in literature. Here we used a symmetric form of the KL-divergence known as Jensen-Shannon divergence (JS-divergence) [27]. By letting $a = \frac{p+q}{2}$ be the average distribution of p and q, the JS-divergence is defined as

$$\mathrm{JS}(p,q) = \frac{1}{2}(\mathrm{KL}(p,a) + \mathrm{KL}(q,a)) = \mathrm{JS}(q,p)$$

and represents the average KL-divergence of the distributions p,q to the average distribution a. When $\mathrm{JS}(p,q)$ is computed in bits we have $0 \le \mathrm{JS}(p,q) \le 1$. Therefore, in this work we always considered the quantity $\log_{10}(\mathrm{JS}(p,q))$.

Experimental procedure

Viable plate counts of *Pseudomonas fluorescens* were obtained essentially as described in [17] in a single time series (six time points). Briefly, bacteria of the rifampicin and tetracyclin resistant strain *P. fluorescens* 92^{RTgfp} carrying the *gfp* gene were precultured to a density of 10^8–10^9 cells/ml. The bacterial suspension was then inoculated into a microcosm consisting of soil supplemented with

biosolid and incubated at 25 C° in the dark over a period of 28 days. Inoculation corresponds to the time point $t=0$. Samples were collected at time points $t=3,7,14,21$ and $t=28$ days, subjected to log10 serial dilution and plated on LB agar added with rifampicin, tetracycline and cycloheximide. For each time point, viable plate counts were determined from five or more technical replicates.

Estimation of survival curves

For each time point, posterior distributions were computed using dupiR and sampling without replacement. Survival curves were fit using a power-law model

$$\log_{10}(n(t)+1)=\frac{\alpha}{(t-t_0)^{\beta}}$$

where $n(t)$ is the maximum a posteriori of the population size at time point t. The model was fit using the R function nls [12] and starting estimates $\alpha=\log_{10}(n_0+1)$, $\beta=0.2$ and $t_0=-0.1$.

Supporting Information

Text S1 Appendix. This supplementary file is an Appendix containing the proof of Lemma 1 and additional information pertaining the derivation of the posterior distributions discussed in the main text.
(PDF)

Figure S1 Simulation results. JS-divergence (expressed in $\log_{10}$) between posterior distributions computed with our method without replacement or with the GP method as a function of the total sampling fractions (R). Total counts $K\in\{0,1,\ldots,30\}$ have been considered.
(TIF)

Figure S2 Maximum JS-divergence as a function of the total sampling fraction. Maximum JS-divergence as a function of R. The red line indicates a linear regression fit. The orange vertical dashed line indicates the value $R=1/32$. For $R<1/32$, posterior distributions computed with or without replacement can be considered to be the same for any practical application.
(TIF)

Figure S3 Comparison between dupiR and the GP method for different Gamma prior parameters. JS-divergence (expressed in $\log_{10}$) between posterior distributions computed with dupiR and sampling without replacement or with the GP method (Clough *et al.* [14]) as a function of total counts (K) and total sampling fractions (R) obtained from two measurements ($m=2$, see Methods). The rate parameter (ρ) of the Gamma prior was varied over four orders of magnitude and different panels correspond to simulations run with (A) $\rho=10^{-5}$, (B) $\rho=10^{-4}$, (C) $\rho=10^{-3}$, (D) $\rho=10^{-2}$, (E) $\rho=0.1$. (F) Example of the effect of the Gamma prior parametrization on the posterior distribution inferred from $K=5, R=0.25$ and $\rho=10^{-6}$ (orange), $\rho=10^{-2}$ (red) and $\rho=0.1$ (brown). The latter case encodes a prior of $K=1$ from $R=0.1$. The posterior distribution estimated with dupiR is shown in blue, with the maximum a posteriori of n indicated by the dashed gray line.
(TIF)

Acknowledgments

F.C. is part of the Life Science Zurich Graduate School, PhD program in Systems Biology. F.C. would like to thank Renato Paro for having kindly supported this work.

Author Contributions

Conceived and designed the experiments: FC MR. Performed the experiments: LF. Analyzed the data: FC LF MR. Wrote the paper: FC MR.

References

1. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: 1–12.
2. Robinson M, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23: 2881–2887.
3. Robinson M, Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 23: 321–332.
4. Morita S, Thall PF, Müller P (2008) Determining the effective sample size of a parametric prior. Biometrics 64: 595–602.
5. Morita S, Thall PF, Müller P (2010) Evaluating the Impact of Prior Assumptions in Bayesian Biostatistics. Stat Biosci 2: 1–17.
6. Bolstad WM (2007) Introduction to Bayesian Statistics, 2nd edition. Wiley-Interscience, 464p.
7. Keynes JM (1921) A treatise on probability. London: Macmillan & Co. 500p.
8. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, 1067p.
9. Jaynes ET (1968) Prior Probabilities. IEEE Trans Syst Sci Cybern, 4: 227–241.
10. Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press. 758p.
11. Cover TM, Thomas AJ (2006) Elements of Information Theory (2nd edition). Wiley-Interscience. 776p.
12. R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
13. Hazel MD (2011) Life, Death, and In-Between: Meanings and Methods in Microbiology. Appl Environ Microbiol 77(16): 5571–5576.
14. Clough HE, Clancy D, O'Neill PD, Robinson SE, French NP (2005) Quantifying Uncertainty Associated with Microbial Count Data: A Bayesian Approach. Biometrics 61: 610–616.
15. Niemela SI (2003) Measurement uncertainty of microbiological viable counts, Accred Qual Assur 8: 559–563.
16. Koch AL (1994) Growth measurement. In: Gerhardt PZ editor. Methods for General and Molecular Bacteriology. ASM Press, Washington DC.
17. Fracchia L, Perotti EBR, Pidello A, Rinaldi M, Martinotti MG (2011) Persistence and Impact of a PGPR on Microbial Communities of Biosolids and Soil Amended with Them, J Environ Eng Sci 5: 578–595.
18. Zwietering MH, Jongenburger FM, Rombouts FM, Van 'Triet K (1990) Modeling of the Bacterial Growth Curve. Appl Environ Microbiol 56(6): 1875–1881.
19. Xiong R, Xie G, Edmondson AE, Sheard MA (1999) A mathematical model for bacterial inactivation. Int J Food Microbiol 46: 45–55.
20. Bates DM, Watts DG (1988) Nonlinear Regression Analysis and Its Applications. Wiley- Interscience. 365p.
21. Bates DM, Chambers JM (1992) Nonlinear models. In: Chambers JM, Hastie TJ editors. Statistical Models in S, Wadsworth & Brooks/Cole.
22. Jaynes ET (1973) The Well-Posed Problem. Found Phys 3: 477–492.
23. Kirchman D, Sigda J, Kapuscinski R, Mitchell R (1982) Statistical analysis of the direct count method for enumerating bacteria. Appl Environ Microbiol 44(2): 376–382.
24. Herigstad B, Hamilton M., Heersink J (2001) How to optimize the drop plate method for enumerating bacteria. J Microbiol Methods 44(2): 121–129.
25. Brugger SD, Baumberger C, Jost M, Jenni W, Brugger U, et al. (2012) Automated Counting of Bacterial Colony Forming Units on Agar Plates. PLoS ONE 7(3): e33695.
26. Kullback S, Leibler RA (1951) On Information and Sufficiency. Ann Math Stat 33: 79–86.
27. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 37: 145–151.