Stochastic Environmental Research and Risk Assessment

Sample size for estimating organism concentration in ballast water: a Bayesian approach

--Manuscript Draft--

Manuscript Number:	
Full Title:	Sample size for estimating organism concentration in ballast water: a Bayesian approach
Article Type:	Original research
Keywords:	sample size; average coverage criterion; average length criterion; Poisson distribution; negative binomial distribution
Corresponding Author:	Eliardo G Costa, Ph.D. Universidade Federal do Rio Grande do Norte Natal, Rio Grande do Norte BRAZIL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universidade Federal do Rio Grande do Norte
Corresponding Author's Secondary Institution:	
First Author:	Eliardo G Costa, Ph.D.
First Author Secondary Information:	
Order of Authors:	Eliardo G Costa, Ph.D. Carlos Daniel Paulino Julio M Singer
Order of Authors Secondary Information:	
Funding Information:	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (No number) — Dr. Eliardo G Costa Conselho Nacional de Desenvolvimento Científico e Tecnológico (BR) (153526/2014-9) — Dr. Eliardo G Costa Conselho Nacional de Desenvolvimento Científico e Tecnológico (3304126/2015-2) — Dr. Julio M Singer Fundação de Amparo à Pesquisa do Estado de São Paulo (2013/21728-2) — Dr. Julio M Singer
Abstract:	Estimation of microorganism concentration in ballast water tanks is important to evaluate and possibly to prevent the introduction of invasive species in stable ecosystems. For such purpose, the number of organisms in ballast water aliquots are counted and used to estimate their concentration with some precision requirement. Poisson and negative binomial models have been employed in this context under a frequentist perspective, the former being appropriate when the organism distribution in the tank is homogeneous and the latter when the organisms are heterogeneously distributed. A Bayesian approach is a flexible alternative since it naturally provides a sequential way of enhancing the estimation procedure by updating the prior distribution along the ballast water discharging process. We adopt such an approach by considering a gamma prior distribution for the mean of the Poisson model and a Pearson type VI prior distribution for the corresponding parameter of the negative binomial model. We propose algorithms to obtain minimum sample sizes required to construct highest posterior density (HPD) credible intervals satisfying average coverage and average length criteria. We also conduct a simulation study to verify whether HPD intervals constructed under either model satisfy the proposed criteria.

Suggested Reviewers:	Lurdes Inoue University of Washington linoue@uw.edu Researcher with papers on sample size determination in a Bayesian approach
	Lawrence Pettit Queen Mary University of London l.pettit@qmul.ac.uk Researcher with papers on sample size determination in a Bayesian approach
	Cyr M'Lan University of Connecticut mlan@merlot.stat.uconn.edu Researcher with papers on sample size determination in a Bayesian approach

Natal, 15th April, 2019

Professor George Christakos

Editor-in-Chief, Stochastic Environmental Research and Risk Assessment

Dear Professor Christakos:

Please find enclosed a copy of the manuscript entitled "Sample size for estimating organism concentration in ballast water: a Bayesian approach" by Costa, E.G., Paulino, C.D. and Singer, J.M. submitted for consideration of possible publication in SERRA. The paper deals with sampling ballast water to verify compliance with the D-2 standards of the International Maritime Organization, a problem that still has not been adequately solved and is the object of ongoing research. We provide methodologies to compute sample sizes to estimate the concentration of organisms in ballast water under a Bayesian approach. We believe that SERRA is the appropriate home for our work and we are looking forward for your comments, suggestions and decision.

Sincerely,

Eliardo G. Costa (corresponding author)

Departamento de Estatística,

Universidade Federal do Rio Grande do Norte

eliardocosta@ccet.ufrn.br

Sample size for estimating organism concentration in ballast water: a Bayesian approach

Eliardo G. Costa*[1], Carlos Daniel Paulino[2], and Julio M. Singer[3]

[1]Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Brazil
[2]Departamento de Matemática, IST and CEAUL, FCUL, Universidade de Lisboa, Portugal
[3]Departamento de Estatística, Universidade de São Paulo, Brazil

Abstract

Estimation of microorganism concentration in ballast water tanks is important to evaluate and possibly to prevent the introduction of invasive species in stable ecosystems. For such purpose, the number of organisms in ballast water aliquots are counted and used to estimate their concentration with some precision requirement. Poisson and negative binomial models have been employed in this context under a frequentist perspective, the former being appropriate when the organism distribution in the tank is homogeneous and the latter when the organisms are heterogeneously distributed. A Bayesian approach is a flexible alternative since it naturally provides a sequential way of enhancing the estimation procedure by updating the prior distribution along the ballast water discharging process. We adopt such an approach by considering a gamma prior distribution for the mean of the Poisson model and a Pearson type VI prior distribution for the corresponding parameter of the negative binomial model. We propose algorithms to obtain minimum sample sizes required to construct highest posterior density (HPD) credible intervals satisfying average coverage and average length criteria. We also conduct a simulation study to verify whether HPD intervals constructed under either model satisfy the proposed criteria.

Keywords: sample size, average coverage criterion, average length criterion, Poisson distribution, negative binomial distribution.

1 Introduction

Evaluation of ballast water discharges from ships is a topic of current interest because the possible introduction of invasive species in stable ecosystems may

*Corresponding author. *E-mail address:* eliardocosta@ccet.ufrn.br

bring serious environmental and economic consequences. Estimates of damage costs of invasive species may vary from 0.4 to 220 (depending on the country) billion USD per year in 2008 prices (Marbuah *et al.*, 2014, Table 1).

The D-2 standard of the International Maritime Organization (IMO) requires that deballasted water should contain no more than 10 living organisms (referred to simply as organisms in the remainder) with maximum dimension 10 μm and 50 μm per mL among other restrictions. Recently, Cohen *et al.* (2017) suggested that the standards must be re-evaluated and the limits must be even smaller. Given the large amount of ballast water (up to thousands of tons) transported by big vessels, one has to rely on sampling methods to verify whether the standard is satisfied. The sampling process is based on a probabilistic model and on a criterion according to which one must compute the number of aliquots of ballast water with volume $w\ mL$ needed to decide whether the D-2 standard is complied with. One of the difficulties with this approach relates to the heterogeneous nature of the organism concentration in the ballast water tank (Murphy *et al.*, 2002; Carney *et al.*, 2013). An overview of research in ballast water in the last thirty years is presented in Bailey (2015).

Based on frequentist methods, Costa *et al.* (2015, 2016) adopted models that take this heterogeneity into account. In particular, Costa *et al.* (2015) consider Poisson and negative binomial distributions and specify probabilities for Type I and II errors to test the hypothesis that the mean organism concentration in the tank is smaller than or equal to 10 organisms per mL. Costa *et al.* (2016), on the other hand, consider the same probability distributions and specify a lower bound to the probability that the difference between the mean concentration and its estimate be less than a fixed value. Although such results are useful, they are essentially dependent on (some) knowledge about the heterogeneity of the organism concentration in the ballast water tank. Costa (2017) suggests the adoption of more flexible models that may possibly incorporate knowledge acquired over time. Bayesian models are excellent candidates to incorporate such characteristics given that such information may be considered in the prior distribution which may also be updated when more data is obtained.

Under a Bayesian approach, two criteria are widely used in the literature for sample size determination, namely, the average coverage and the average length of credible intervals (ACC and ALC, respectively). In both cases, we choose the smallest sample size that satisfies the condition imposed on some specified average characteristic of the posterior distribution of the parameter of interest. For the ACC we compute the posterior probability of a highest posterior density (HPD) interval with fixed length for each sample $\boldsymbol{x}_n$ of size n and weigh it by the marginal distribution of the data. This average probability must be not smaller than a specified lower bound. For the ALC, on the other hand, we compute the length of an interval with fixed credible degree for each $(\boldsymbol{x}_n, n)$ and weigh it by the same marginal distribution. The average length must not be larger than a specified upper bound.

Adcock (1987, 1988) uses ACC (with a different label) to determine sample sizes to estimate multinomial probabilities under Dirichlet prior distributions as well as to estimate the mean and the variance of normal distributions with prior normal or chi-squared distributions for the case where the variance is known or unknown, respectively. Joseph *et al.* (1995) and Joseph *et al.* (1997) use both the ACC and the ALC, among other Bayesian criteria for estimating the proportion and the difference between two proportions under binomial distributions with

beta prior distributions. In problems involving the estimation of the means of normal, binomial and Poisson distributions or of the slope parameter in linear regression models, Adcock (1997) discusses the ACC, the ALC and other Bayesian criteria. Wang & Gelfand (2002) use the same criteria to determine the sample size for the estimation of parameters of distributions belonging to the exponential family, of parameters in Weibull survival models as well as of parameters in logistic regression models. M'Lan *et al.* (2006) use ACC and ALC criteria in the context of case-control studies; Stamey *et al.* (2006) also consider these criteria to estimate the parameters of Poisson distributions as well to estimate the difference or the ratio of the parameters of two Poisson distributions. Nassar *et al.* (2010, 2011) use the same criteria to estimate the parameter of geometric distributions under beta prior distributions and the parameter of Laplace distributions under normal prior distributions. We may also cite Joseph & Bélisle (1997), Joseph & Wolfson (1997), Rahme *et al.* (2000), De Santis (2007), M'Lan *et al.* (2008) for related work. and the corresponding volumes In the context of sample size determination, we consider a Bayesian approach to compute minimum sample sizes required to obtain lower and upper limits of credible intervals for the mean organism concentration in a ballast water tank with specified average coverage or average length. For simulated samples $\boldsymbol{x}_n$ of size n computed according to the proposed approach, the credible intervals defined by the lower [say, $a(\boldsymbol{x}_n)$] and upper [say, $b(\boldsymbol{x}_n)$] limits will have in average, the specified coverage or length. Once the required minimum sample size, say n_{m}, has been determined, a real dataset $\boldsymbol{x}_{n_{\mathrm{m}}}$ will be collected. Then a ship is declared not compliant with the D-2 standard if $a(\boldsymbol{x}_{n_{\mathrm{m}}}) > 10$ or compliant, if $b(\boldsymbol{x}_{n_{\mathrm{m}}}) < 10$. Otherwise, if $a(\boldsymbol{x}_{n_{\mathrm{m}}}) < 10 < b(\boldsymbol{x}_{n_{\mathrm{m}}})$, more data are needed to make a decision.

In Section 2 we describe the adopted Bayesian models. Sample size determination under both the ACC and the ALC criteria is discussed in Section 3. A simulation study to evaluate whether HPD intervals constructed with the proposed sample sizes satisfy the adopted optimality criteria is presented in Section 4. We conclude with a discussion in Section 5. Algorithms for sample size computations, written with the R language (R Core Team, 2016), are presented in the Supplementary Material.

2 Bayesian models

2.1 Poisson model with a gamma prior distribution

Given a mean organism concentration λ, let X be the number of organisms in an aliquot of volume w; in this aliquot, we expect to find $\mathbb{E}[X|\lambda] = w\lambda$ organisms. Suppose that, given λ, X follows a Poisson distribution with mean $w\lambda$, *i.e.*, the organisms are homogeneously distributed in the ballast water tank.

The natural (conjugate) choice for the prior distribution is a gamma distribution with parameters θ_0 and λ_0, namely $\lambda \sim G(\theta_0, \theta_0/\lambda_0)$, for which the probability density function is

$$f(\lambda) \propto \lambda^{\theta_0 - 1} \exp(-\theta_0 \lambda / \lambda_0).$$

This implies that $\mathbb{E}[\lambda] = \lambda_0$ and $\mathrm{Var}[\lambda] = \lambda_0^2/\theta_0$. In this context, λ_0 represents a prior mean concentration and θ_0 controls the variability of λ around λ_0. The

gamma distribution provides ample flexibility to model the shape of the prior knowledge on the mean concentration λ as depicted in Figure 1.

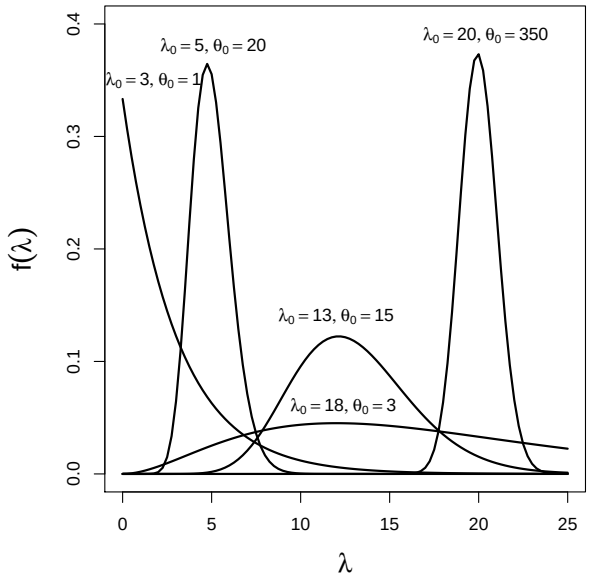

Figure 1: Probability density functions for $G(\theta_0, \theta_0/\lambda_0)$ distributions.

Consider a random sample $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ of size n of $X|\lambda$ and a gamma prior distribution for λ. We may write the Bayesian model hierarchically as follows

$$X_i|\lambda \overset{\text{iid}}{\sim} Poi(w\lambda), \quad i = 1, 2, \ldots, n; \tag{1}$$

$$\lambda \sim G(\theta_0, \theta_0/\lambda_0). \tag{2}$$

In this context, the posterior distribution of λ is also gamma, with parameters $\theta_0 + s_n$ and $nw + \theta_0/\lambda_0$, where $s_n = \sum_{i=1}^{n} x_i$, *i.e.*, $\lambda|\boldsymbol{x}_n \sim G(\theta_0 + s_n, nw + \theta_0/\lambda_0)$.

An example of prior and posterior densities is presented in Figure 2. The effect of the observed data is clearly observed to lead to a posterior distribution more concentrated than the prior distribution.

2.2 Negative binomial model with a Pearson Type VI prior distribution

In contrast with the homogeneity assumption for the organism distribution in the tank inherent to the Poisson model, consider a more realistic situation where the organisms are distributed heterogeneously.

Assume that the organism concentration in the i-th aliquot is λ_i and the corresponding number of organisms is X_i, $i = 1, \ldots, n$. Then, in the i-th aliquot we expect to find $\mathbb{E}[X_i|\lambda_i] = w\lambda_i$ organisms. For $i = 1, \ldots, n$, suppose that, given λ_i, X_i follows a Poisson distribution with mean $w\lambda_i$ and that given λ and ϕ, $\lambda_i \sim G(\phi, \phi/\lambda)$ so that $\mathbb{E}[\lambda_i|\lambda] = \lambda$ and $\text{Var}[\lambda_i|\lambda] = \lambda^2/\phi$. Thus, given λ

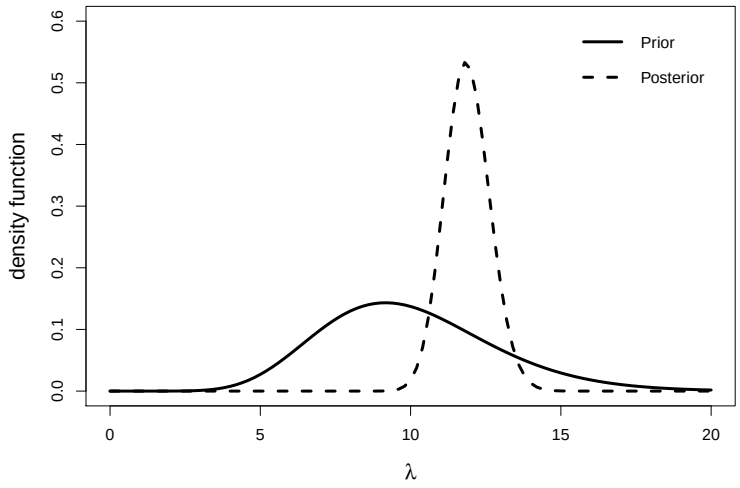

Figure 2: Prior gamma distribution $G(\theta_0, \theta/\lambda_0)$ and posterior gamma distribution $G(\theta_0 + s_n, nw + \theta_0/\lambda_0)$ with $w = 1$, $\lambda_0 = 10$, $\theta_0 = 12$, $n = 20$ and $s_n = 240$.

and ϕ, X_i follows a negative binomial distribution with $\mathbb{E}[X_i|\lambda, \phi] = w\lambda$ and $\text{Var}[X_i|\lambda, \phi] = w\lambda + (w\lambda)^2/\phi$, where ϕ is a shape (or agglomeration) parameter, assumed known. This is denoted as $X_i|\lambda, \phi \sim NB(w\lambda, \phi)$.

A natural conjugate prior distribution for the negative binomial distribution is the Pearson Type VI distribution, also known as the beta prime distribution (Johnson *et al.*, 1994a,b), for which the kernel of the probability density function is

$$f(\lambda) \propto \left(\frac{w}{\phi}\lambda\right)^{\theta_0 - 1} \left(1 + \frac{w}{\phi}\lambda\right)^{-\theta_0 - (\theta_0/\lambda_0 + 1)},$$

with location parameter 0, scale parameter ϕ/w and shape parameters θ_0 and $\theta_0/\lambda_0 + 1$, where λ_0 and θ_0 are known positive fixed constants (hyperparameters). We use the notation $\lambda \sim PVI(0, \phi/w, \theta_0, \theta_0/\lambda_0 + 1)$. In this case, $\mathbb{E}[\lambda] = (\phi/w)\lambda_0$ and $\text{Var}[\lambda] = (\lambda_0^2/\theta_0)[\phi^2(\lambda_0 + 1)/(w^2(1 - \lambda_0/\theta_0))]$, for $\lambda_0 < \theta_0$.

In the Poisson model with gamma prior distribution, we have $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\lambda]] = \mathbb{E}[w\lambda] = w\lambda_0$, *i.e.*, the expected number of organisms when collecting a aliquot depends only on the hyperparameter λ_0. This makes sense since we are assuming homogeneity for the concentration, and regardless of the location where we collect an aliquot in the ballast water tank, we expect to find the same number of organisms. On the other hand, if we consider the negative binomial model with a Pearson Type VI prior distribution, we have $\mathbb{E}[X] = \phi\lambda_0$ so that the expected number of organisms in an aliquot depends on the parameter ϕ that controls the heterogeneity of the organisms in the tank. Note that ϕ is also a scale parameter for the prior distribution and the larger its value, the more spread out is the distribution with the other parameters fixed, indicating a vague prior knowledge about the parameter of interest. Furthermore, we can set the other parameters in such a way that the prior distribution may represent

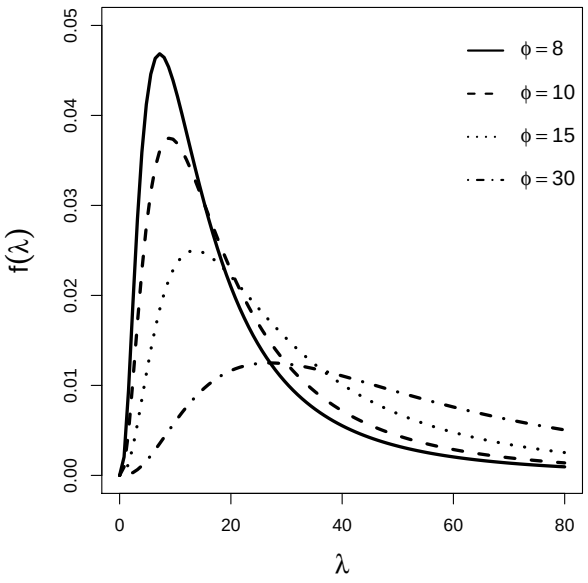

Figure 3: Probability density functions for $PVI(0, \phi/w, \theta_0, \theta_0/\lambda_0 + 1)$ distributions with $w = 1$, $\lambda_0 = 3$ and $\theta_0 = 4$.

cases where there is high probability associated to an interval even when the value of ϕ increases (Figure 4). When λ_0 and θ_0 are fixed and ϕ increases, we have distributions representing cases with large variability (Figure 3).

Consider a random sample of size n from $X|\lambda, \phi$, and a Pearson Type VI prior distribution for λ. We may write the model hierarchically as follows

$$X_i|\lambda, \phi \overset{\text{iid}}{\sim} NB(w\lambda, \phi), \quad i = 1, 2, \ldots, n; \tag{3}$$

$$\lambda \sim PVI(0, \phi/w, \theta_0, \theta_0/\lambda_0 + 1). \tag{4}$$

In this context, the posterior distribution of λ is Pearson Type VI, with the same location and scale parameters as the prior distribution, and shape parameters $\theta_0 + s_n$ and $\theta_0/\lambda_0 + n\phi + 1$, *i.e.*, $\lambda|\boldsymbol{x}_n \sim PVI(0, \phi/w, \theta_0 + s_n, \theta_0/\lambda_0 + n\phi + 1)$.

We must emphasize that ϕ plays two roles in model (3)-(4). In (3), it plays the role of a dispersion (or agglomeration) parameter. The larger is ϕ, the more homogeneous is the organism concentration in the tank. In the prior distribution (4), ϕ plays the role of scale parameter. Keeping the other parameters fixed, the larger is ϕ, the less precise is the prior knowledge about the parameter of interest (see Figure 3). This does not mean that if ϕ (previously known) is large we may only assign prior distributions with large variability, because we may specify the parameters λ_0 and θ_0 to adjust the precision of the prior knowledge even with large values of ϕ (see Figure 4).

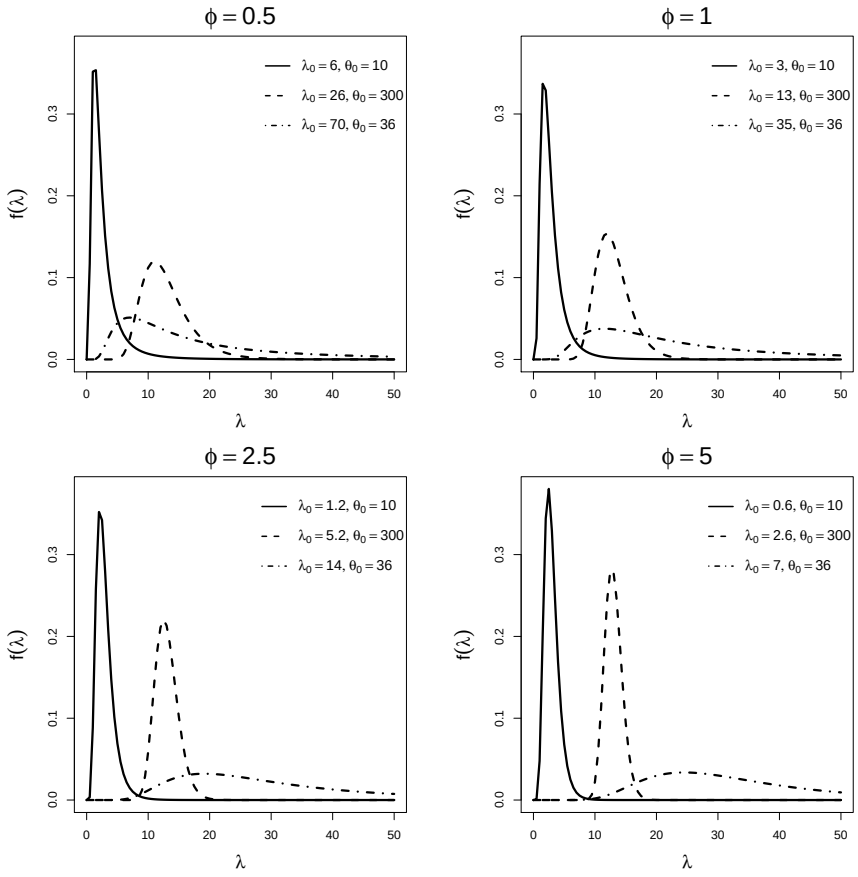

Figure 4: Probability density functions of $PVI(0, \phi/w, \theta_0, \theta_0/\lambda_0+1)$ distributions with $w = 1$.

3 Sample size determination

We consider two criteria to determine the minimum sample size required to estimate λ with a pre-specified precision.

3.1 Average coverage criterion (ACC)

The objective is to obtain the minimum sample size n such that the credible interval $R(\boldsymbol{x}_n)$ for λ has a pre-specified length with posterior probability at least equal to $1-\rho$, *i.e.*,

$$\int_{R(\boldsymbol{x}_n)} f(\lambda|\boldsymbol{x}_n)d\lambda \geq 1-\rho,$$

where $\boldsymbol{x}_n$ is a sample of size n and $R(\boldsymbol{x}_n)$ is a subset (an interval in our case) of the parameter space. Since the sample size determination precedes the actual sampling, we must consider all possible outcomes for $\boldsymbol{x}_n$ to achieve the objective.

In this direction we may weigh each outcome by its probability, *i.e.*,

$$\int_{\mathcal{X}^n}\left[\int_{R(\boldsymbol{x}_n)} f(\lambda|\boldsymbol{x}_n)d\lambda\right] g(\boldsymbol{x}_n)d\boldsymbol{x}_n \geq 1-\rho,$$

where $\mathcal{X}^n$ is the sample space associated to $\boldsymbol{x}_n$ and $g(\boldsymbol{x}_n)$ is the marginal probability (or density) function of the outcomes.

For models (1)-(2) and (3)-(4), the credible region may be an interval and in this case we consider the highest posterior density (HPD) interval to define $R(\boldsymbol{x}_n) = [a(\boldsymbol{x}_n), b(\boldsymbol{x}_n)]$. We fix the length $\ell > 0$ of the HPD intervals for λ, specify the minimum Bayesian coverage probability, $1-\rho$ and determine the minimum sample size as well as the bounds $a(\boldsymbol{x}_n)$ and $b(\boldsymbol{x}_n) = a(\boldsymbol{x}_n) + \ell$ such that

$$\int_{\mathcal{X}^n}\left[\int_{a(\boldsymbol{x}_n)}^{b(\boldsymbol{x}_n)} f(\lambda|\boldsymbol{x}_n)d\lambda\right] g(\boldsymbol{x}_n)d\boldsymbol{x}_n \geq 1-\rho. \tag{5}$$

Given $a(\boldsymbol{x}_n)$, $b(\boldsymbol{x}_n)$ and the parameters of the posterior distribution, the inner integral in (5) may be obtained computationally; the outer integral may be estimated via Monte Carlo simulation. An algorithm to obtain the minimum sample size satisfying the criterion is outlined in the Supplementary Material.

In Tables 1 and 2 we present sample sizes computed using ACC (5) for models (1)-(2) and (3)-(4). Note that in the case of model (3)-(4) we consider $\lambda_0 = 10(w/\phi)$ to make the prior expected value equal to 10 in order to allow a comparison with the model (1)-(2) for which we fixed $\lambda_0 = 10$.

Table 1: ACC (5) based minimum sample size (n) computed under the Poisson/gamma model (1)-(2) with $\lambda_0 = 10$ and $\rho = 0.05$.

Aliquot volume (w)	Interval length (ℓ)	Shape parameter (θ_0)				
		1.0	2.5	5.0	7.5	10.0
0.5	2	77	77	76	76	75
	4	20	19	19	18	18
1.0	2	39	39	38	38	38
	4	10	10	10	9	9

For illustrative purposes, we obtain HPD intervals based on a set of hypothetical counts generated to mimic real heterogeneously distributed data. We first determined the sample size required to satisfy the ACC with $\ell = 2$ assuming a Poisson/gamma model with a prior distribution having $\lambda_0 = 10$ and $\theta_0 = 0.01$ in order to obtain a large variance. Setting $w = 1$, the required sample size is $n_{\mathrm{P}} = 59$. We then generated 59 observations via a negative binomial model with $\lambda = 9$, $\phi = 0.1$ and $w = 1$. The generated counts are displayed in Table 3 where the heterogeneity induced by the negative binomial/Pearson Type VI model is evident.

The sum of the counts in Table 3 is $s_{n_{\mathrm{P}}} = s_{59} = 653$ so that the corresponding HPD intervals (obtained via the algorithms described described in Subsections 1.1.1 and 1.2.1 of the Supplementary Material) are, respectively, (10.08, 12.08) for the Poisson/gamma model, and (7.11, 9.11) for the negative binomial/Pearson Type VI model (in this case, we set $\phi = 0.0899$, obtained via maximum likelihood).

Table 2: ACC (5) based minimum sample size (n) computed under the negative binomial/Pearson Type VI model (3)-(4) with $\lambda_0 = 10(w/\phi)$ and $\rho = 0.05$.

Aliquot volume (w)	Interval length (ℓ)	ϕ	Shape parameter (θ_0)			
			11	25	50	75
0.5	2	1.0	462	457	453	444
		2.5	229	226	222	216
		5.0	152	149	144	140
		7.5	127	124	118	114
		10.0	113	111	106	101
	4	1.0	115	112	106	101
		2.5	57	53	49	43
		5.0	37	34	29	24
		7.5	30	28	22	17
		10.0	27	24	19	14
1.0	2	1.0	426	422	417	414
		2.5	194	191	188	185
		5.0	115	113	111	108
		7.5	90	88	85	82
		10.0	77	75	72	70
	4	1.0	110	107	102	99
		2.5	49	46	44	41
		5.0	29	27	24	22
		7.5	22	20	18	15
		10.0	19	17	15	12

Table 3: Simulated counts for the example under the negative binomial model with $\phi = 0.1$, $\lambda = 9$ and $w = 1$.

0	0	0	3	0	20	0	0	29	4	2	10	3	0	0
97	0	39	0	0	1	0	0	0	0	0	0	0	0	313
0	0	0	3	0	1	1	6	0	0	13	0	0	0	0
18	0	5	0	0	0	5	0	4	0	0	76	0	0	

The first interval does not contain the organism concentration $\lambda = 9$ and suggest non-compliance with the D-2 regulation. The second interval, on the other hand, contains $\lambda = 9$ (even with the sample size obtained under the Poisson/gamma model) and suggests compliance with the D-2 regulation.

3.2 Average length criterion (ALC)

An alternative criterion used to determine sample sizes is based on the average length of the posterior credible intervals. The rationale here is to set the minimum Bayesian coverage probability $1 - \rho$ and obtain the minimum sample size n by requiring that the length of the posterior credible region $\ell'(\boldsymbol{x}_n, n) = b(\boldsymbol{x}_n) - a(\boldsymbol{x}_n)$ be such that

$$\int_{\mathcal{X}^n} \ell'(\boldsymbol{x}_n, n) g(\boldsymbol{x}_n) d\boldsymbol{x}_n \leq \ell_{\max}, \tag{6}$$

where $\ell_{\max}$ is the maximum admissible length for the posterior credible region.

The lower and upper bounds of the HPD interval may be obtained via numerical methods and the integral by Monte Carlo simulation. An algorithm to obtain the minimum sample size satisfying this criterion is outlined in the Supplementary Material.

Based on the ideas of M'Lan *et al.* (2008), who used a binomial model with a beta prior distribution, we may obtain the sample size using the ALC under the model (1)-(2) with no need for numerical methods via the following result.

Theorem 1 *Consider the Poisson/gamma (1)-(2) model and the average length criterion (6). The minimum n, based on large sample approximation, to guarantee that the posterior credible interval average length is smaller than $\ell_{\max}$ is the smallest integer such that*

$$n \geq \frac{\theta_0}{w\lambda_0}\left\{\left[\frac{\lambda_0}{\theta_0}\frac{2z_{\rho/2}}{\ell_{\max}}\frac{\Gamma(\theta_0+1/2)}{\Gamma(\theta_0)}\right]^2 - 1\right\},$$

where $z_{\rho/2}$ is the quantile of order $1-\rho/2$ of the standard normal distribution.

The proof of Theorem 1 is presented in the Supplementary Material. In Tables 4 and 5 we present sample sizes computed using ALC (6) for models (1)-(2) and (3)-(4); in Table 4 we present corresponding sample sizes (within parentheses) computed using Theorem 1.

Table 4: ALC (6) based minimum sample size (n) computed under the Poisson/gamma model (1)-(2) (and also using Theorem 1) with $\lambda_0 = 10$ and $\rho = 0.05$.

Aliquot volume (w)	Maximum interval length ($\ell_{\max}$)	Shape parameter (θ_0)				
		1.0	2.5	5.0	7.5	10.0
0.5	2	77 (61)	77 (70)	76 (73)	76 (73)	75 (73)
	4	19 (15)	19 (17)	19 (18)	18 (18)	17 (17)
1.0	2	38 (31)	38 (35)	38 (37)	38 (37)	38 (37)
	4	10 (8)	10 (9)	9 (9)	9 (9)	9 (9)

4 Simulation study

For each (prior distribution) scenario and sample size obtained via the ACC (5) displayed in Table 1 we drew 1000 samples from a Poisson/gamma model (1)-(2) with values of λ fixed at the quantiles of order 1/6, 2/6, 3/6. 4/6 and 5/6 of the corresponding prior distribution. Then, for each sample we obtained the lower $[a(\boldsymbol{x}_n)]$ and upper $[b(\boldsymbol{x}_n)]$ limits of the HPD credible interval for the mean organism concentration in a ballast water tank with pre-specified average coverage probability ($1-\rho = 0.95$) and computed the proportion of intervals containing the fixed value of λ. The results are displayed in Table 6. We expect that the estimates of the HPD Bayesian coverage probability to be at least 0.95.

Under the same model, but using sample sizes displayed in Table 4, obtained via the ALC (6), we conducted a similar simulation study, the results of which are displayed in Table 7. In this case, we expect that the estimates of length of the HPD intervals to be at most 2 or 4.

Table 5: ALC (6) based minimum sample size (n) computed under the negative binomial/Pearson Type VI model (3)-(4) with $\lambda_0 = 10(w/\phi)$ and $\rho = 0.05$.

Aliquot volume (w)	Maximum interval length ($\ell_{\max}$)	ϕ	Shape parameter (θ_0)			
			11	25	50	75
0.5	2	1.0	456	456	452	445
		2.5	228	226	221	216
		5.0	151	150	144	138
		7.5	126	123	118	113
		10.0	114	110	106	101
	4	1.0	113	109	105	99
		2.5	55	52	48	43
		5.0	37	34	29	23
		7.5	30	27	22	17
		10.0	27	24	19	14
1.0	2	1.0	414	418	416	410
		2.5	191	190	186	185
		5.0	115	113	110	108
		7.5	89	88	85	82
		10.0	76	75	72	70
	4	1.0	103	104	100	97
		2.5	47	46	43	41
		5.0	28	26	24	22
		7.5	21	20	18	15
		10.0	19	17	15	12

The same strategy was conducted for data obtained via the negative binomial/Pearson VI model (3)-(4) using the sample sizes provided in Tables 2 and 5. The results are provided in Table 8 and in Tables S2-S4 of the Supplementary Material.

In addition, we applied the same strategy described previously using $\phi = 0.5$ to draw samples $\boldsymbol{x}_n$; we then obtained HPD credible intervals using different values of $\phi > 0.5$ to see how the coverage probabilities $(1 - \rho)$ and HPD interval lengths ($\ell_{\max}$) behave in relation to the specified threshold. The results are displayed in Table 9 and in Tables S5-S7 of the Supplementary Material.

5 Discussion

The results in Table 1 obtained under the Poisson/gamma model indicate that the sample size does not decrease much when θ_0 increases, *i.e.*, when the prior variance decreases. This may be explained by the homogeneity assumption for the concentration which is intrinsic to the adopted model. Unless we consider a precise prior distribution, the sample size required to satisfy the ACC will not change much. This feature is also visible when we compute the sample size under the same model using the ALC (see Table 4).

On the other hand, under model (3)-(4) using either the ACC or the ALC with a fixed value for ϕ, the precision of the prior knowledge, controlled by θ_0 here, directly affects the required sample size. This also happens when we

Table 6: ACC based Bayesian coverage probability of HPD intervals estimated via simulation for some scenarios under the Poisson/gamma model (1)-(2) with sample sizes displayed in Table 1.

Aliquot volume (w)	Interval length (ℓ)	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
0.5	2	1.0	77	1.00	1.00	0.98	0.94	0.85
		2.5	77	0.99	0.99	0.97	0.95	0.88
		5.0	76	0.99	0.98	0.95	0.94	0.92
		7.5	76	0.98	0.97	0.96	0.94	0.91
		10.0	75	0.97	0.98	0.95	0.95	0.93
	4	1.0	20	1.00	1.00	0.99	0.96	0.89
		2.5	19	0.99	0.99	0.97	0.94	0.89
		5.0	19	0.99	0.98	0.97	0.95	0.90
		7.5	18	0.97	0.97	0.97	0.95	0.92
		10.0	18	0.97	0.99	0.97	0.95	0.93
1.0	2	1.0	39	1.00	1.00	0.98	0.95	0.86
		2.5	39	0.99	0.99	0.97	0.93	0.88
		5.0	38	0.99	0.98	0.97	0.94	0.90
		7.5	38	0.98	0.98	0.95	0.94	0.90
		10.0	38	0.98	0.96	0.95	0.96	0.92
	4	1.0	10	1.00	1.00	0.98	0.95	0.88
		2.5	10	0.99	0.99	0.97	0.95	0.89
		5.0	10	0.99	0.98	0.97	0.95	0.92
		7.5	9	0.98	0.97	0.97	0.96	0.92
		10.0	9	0.97	0.97	0.97	0.96	0.91

consider a fixed θ_0 and vary ϕ, that plays the role of a scale parameter in the prior distribution (see Tables 2 and 5).

The convenience of assuming that ϕ is known is a disadvantage but we may circumvent this problem in a practical manner without considering a prior distribution for this parameter. The first and simpler way is to consider ϕ as small as possible, *e.g.*, $\phi = 0.5$. Since the sample size n decreases as ϕ increases, when we take ϕ as the minimum, we are being conservative, in the sense that the corresponding n is enough or more than enough to achieve the pre-specified criteria settings. The results in Table 9 indicate that in some cases the choice of the parameter ϕ has a negligible effect in the computation of the minimum sample size, n. For example, for $\ell = 2$ and θ_0 at most 25, the minimum coverage probability is satisfied for $\phi \leq 10$, indicating that we may take $n = 111$. If we consider θ_0 at most 50, the minimum coverage probability is satisfied for $\phi \leq 2$, indicating that we may take $n = 222$. The same behavior holds for $w = 1$ under either the ACC and or the ALC as indicated in Tables S5-S7 of the Supplementary Material.

The second alternative is to consider a naive sequential procedure in which samples are selected one by one (or by lots). Observe that sample sizes obtained under a Poisson/gamma model (n_P) are always smaller than those obtained by a negative binomial/Pearson VI model (n_{NB}), with respective parameters fixed and write $n_{\mathrm{NB}} = n_{\mathrm{P}} + K$, where K is a positive integer. For fixed w, ℓ (or

Table 7: ALC based length of HPD intervals estimated via simulation for some scenarios under the Poisson/gamma model (1)-(2) with sample sizes displayed in Table 4.

Aliquot volume (w)	Maximum interval length ($\ell_{\max}$)	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
0.5	2	1.0	77	0.85	1.27	1.66	2.09	2.66
		2.5	77	1.30	1.60	1.86	2.13	2.49
		5.0	76	1.52	1.75	1.93	2.12	2.37
		7.5	76	1.61	1.79	1.94	2.10	2.30
		10.0	75	1.67	1.83	1.96	2.10	2.27
	4	1.0	19	1.70	2.54	3.33	4.18	5.33
		2.5	19	2.59	3.20	3.70	4.23	4.94
		5.0	19	3.00	3.43	3.78	4.15	4.62
		7.5	18	3.24	3.58	3.86	4.18	4.55
		10.0	17	3.40	3.71	3.95	4.20	4.52
1.0	2	1.0	38	0.86	1.28	1.67	2.10	2.68
		2.5	38	1.31	1.61	1.87	2.14	2.50
		5.0	38	1.52	1.74	1.93	2.12	2.37
		7.5	38	1.61	1.80	1.95	2.10	2.30
		10.0	38	1.66	1.82	1.95	2.08	2.26
	4	1.0	10	1.67	2.48	3.25	4.07	5.21
		2.5	10	2.53	3.10	3.61	4.12	4.82
		5.0	9	3.09	3.52	3.88	4.25	4.75
		7.5	9	3.24	3.58	3.88	4.17	4.54
		10.0	9	3.32	3.60	3.85	4.09	4.42

$\ell_{\max}$) and hyperparameters, we may compute the sample size under a Poisson model, proceed with the sample collection obtaining n_{P} organism counts ($\boldsymbol{x}_{n_{\mathrm{P}}}$). Using these n_{P} organism counts we may compute an estimate for ϕ by maximum likelihood or by the method of moments (see Ludwig & Reynolds, 1988, eq. 3.5, for example) and with this estimate we may obtain n_{NB} and consequently K, which is the required number of additional aliquots. Since the prior distributions used in both models are different, we must choose the hyperparameters for the Pearson Type VI distribution which represent "equivalent prior knowledge" to those fixed in the gamma distribution. Given w, λ_0 and the estimate of ϕ, we may choose θ_0 such that the plot of the Pearson Type VI distribution is similar to the plot of the gamma distribution with previous hyperparameters used to obtain n_{P}.

The standard approach would be to consider a prior distribution for ϕ which implies setting at least an additional hyperparameter so that we must deal with another integral in order to obtain the marginal distribution of λ. This introduces further computational effort and is the object of future research.

As in Inoue *et al.* (2005), we compare sample sizes obtained under different perspectives. Under the Bayesian approach fixing either ℓ or $\ell_{\max}$ (Tables 2 and 5) the sample sizes are, in general, smaller than those computed under a frequentist approach with ϵ_a (maximum absolute error estimation) equal to 1 or 2

Table 8: ACC based Bayesian coverage probability for HPD intervals estimated via simulation for some scenarios under the negative binomial/Pearson VI model (3)-(4) with sample sizes displayed in Table 2 setting $w = 0.5$.

Interval length (ℓ)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	462	1.00	1.00	0.99	0.95	0.83
		25	457	1.00	1.00	0.98	0.94	0.85
		50	453	0.99	0.98	0.95	0.94	0.88
		75	444	0.99	0.98	0.96	0.94	0.89
	2.5	11	229	1.00	0.99	0.98	0.94	0.87
		25	226	0.99	0.98	0.96	0.93	0.87
		50	222	0.98	0.97	0.97	0.95	0.90
		75	216	0.97	0.97	0.96	0.94	0.92
	5.0	11	152	0.99	0.98	0.96	0.94	0.87
		25	149	0.98	0.97	0.95	0.92	0.92
		50	144	0.96	0.97	0.97	0.95	0.93
		75	140	0.96	0.97	0.96	0.96	0.93
	7.5	11	127	0.99	0.98	0.96	0.94	0.89
		25	124	0.98	0.97	0.96	0.94	0.91
		50	118	0.97	0.97	0.96	0.95	0.93
		75	114	0.97	0.98	0.96	0.95	0.94
	10.0	11	113	0.98	0.99	0.96	0.93	0.89
		25	111	0.98	0.97	0.97	0.94	0.93
		50	106	0.96	0.96	0.96	0.94	0.93
		75	101	0.96	0.97	0.97	0.95	0.95
4	1.0	11	115	1.00	1.00	0.99	0.95	0.84
		25	112	1.00	0.99	0.98	0.96	0.86
		50	106	0.99	0.98	0.97	0.97	0.91
		75	101	0.98	0.97	0.97	0.95	0.91
	2.5	11	57	0.99	0.98	0.99	0.94	0.85
		25	53	0.98	0.98	0.97	0.95	0.90
		50	49	0.97	0.99	0.98	0.97	0.93
		75	43	0.96	0.98	0.98	0.99	0.94
	5.0	11	37	0.99	0.99	0.98	0.95	0.90
		25	34	0.98	0.98	0.97	0.95	0.90
		50	29	0.97	0.98	0.98	0.98	0.94
		75	24	0.97	0.99	0.99	0.99	0.95
	7.5	11	30	0.98	0.99	0.98	0.94	0.90
		25	28	0.98	0.98	0.97	0.96	0.93
		50	22	0.96	0.98	0.99	0.97	0.95
		75	17	0.96	0.99	1.00	0.99	0.96
	10.0	11	27	0.98	0.98	0.97	0.95	0.91
		25	24	0.98	0.97	0.97	0.95	0.93
		50	19	0.97	0.99	0.99	0.98	0.95
		75	14	0.97	0.99	1.00	0.99	0.99

(see Tables 2 and 3 in Costa *et al.*, 2016). This may be justified by the additional information provided by the prior distribution relatively to that considered in the frequentist approach, where only lower and upper bounds for the parameter of interest are given.

For the ALC we present a result (Theorem 1) which allows the computation

of sample sizes under model (1)-(2) without the need for numerical and/or simulation methods. The corresponding sample sizes are consistently smaller than those obtained via Monte Carlo replicates, although the differences are not large. Note that since this theorem is based on large sample approximations, we expect a difference between the corresponding sample sizes and those obtained directly from the proposed criterion.

The simulation results (Tables 6-8 and Tables S2-S4 of the Supplementary Material) show sample sizes similar to those obtained under the simulation study presented in Costa *et al.* (2016). For smaller values of λ, the coverage criterion is attained well above the limit but the results are reversed for the larger values and the minimum fixed coverage is not attained. A similar conclusion holds when using the ALC. We also note that for values of λ smaller or equal to the median, the estimated coverage probability is larger than the proposed one. This is expected, but may not happen for values of λ greater than the median, mainly for the quantiles of order 5/6 or higher, *i.e.*, in some cases the posterior interval does not contain λ, and this happens with estimated coverage probability smaller than the specified one. This suggests that in a practical situation if we want a minimum coverage with probability $1 - \rho$, we should consider a sample size n corresponding to a minimum coverage probability greater than $1 - \rho$ in order to prevent or minimize this problem.

Assuming a parametric distribution for the prior distribution may not provide a realistic picture of the organism distribution in the ballast water tank, especially given the lack of observational or experimental data. For example, we may imagine a situation where two different regions of the tank have large organism concentration, while other regions have not. An alternative is to consider a nonparametric approach where the form of the distribution of the organism concentration is not specified. An approach based on a Dirichlet process mixture as in (Ferguson, 1973; Antoniak, 1974) is currently under investigation.

Practical issues related to the actual collection of the ballast water aliquots have been addressed by many authors (Carney *et al.*, 2013; First *et al.*, 2013; Gollasch & David, 2017). Among them we mention the difficulty in accessing the ballast water tank and the need to submit the sampled aliquots for analysis in a laboratory. Therefore, some of the proposed sample sizes (*e.g.*, 462 in Table 2) are unrealistic with the present technology. However, researchers at the Oceanographic Institute of the University of São Paulo are developing a system in which part of the discharged water will be conducted through an optical device where the organisms will be counted by an appropriate software. This will allow the collection of a very large number of aliquots along the entire deballasting process. We intend to feed counts acquired according to the proposed sample sizes to a computer where the the mean concentration may be estimated. A prototype of the equipment being developed was used for other purposes as indicated in (Matuszewski *et al.*, 2015). Unfortunately we still do not have experimental data obtained via the system being developed.

Although the focus of this study is ballast water sampling, similar results may be applied to other problems in which the Poisson or the negative binomial models underlie the data generating process. For additional applications in the context of Biostatistics see Ludwig & Reynolds (1988) and White & Bennetts (1996), for example.

Table 9: ACC based Bayesian coverage probability for HPD intervals estimated via simulation with $\phi = 0.5$ and computed with different values $\phi > 0.5$ under the negative binomila/Pearson VI model (3)-(4) with sample obtained from Table 2 with $w = 0.5$.

Interval length (ℓ)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	462	1.000	0.999	0.999	0.992	0.940
		25	457	1.000	0.997	0.996	0.990	0.957
		50	453	0.997	0.995	0.989	0.987	0.963
		75	444	0.993	0.987	0.984	0.983	0.954
	2.5	11	229	1.000	1.000	0.999	1.000	0.994
		25	226	1.000	0.999	0.998	0.996	0.985
		50	222	0.989	0.990	0.989	0.984	0.981
		75	216	0.948	0.935	0.934	0.926	0.923
	5.0	11	152	1.000	1.000	1.000	1.000	0.999
		25	149	0.998	0.999	0.999	0.999	0.997
		50	144	0.905	0.924	0.934	0.930	0.922
		75	140	0.236	0.326	0.348	0.407	0.436
	7.5	11	127	1.000	1.000	1.000	1.000	0.999
		25	124	1.000	0.998	0.997	0.998	0.994
		50	118	0.581	0.626	0.647	0.680	0.720
		75	114	0.000	0.001	0.000	0.006	0.018
	10.0	11	113	1.000	1.000	1.000	1.000	1.000
		25	111	0.997	0.998	0.995	0.999	0.997
		50	106	0.163	0.192	0.236	0.305	0.379
		75	101	0.000	0.000	0.000	0.000	0.000
4	1.0	11	115	0.999	1.000	0.993	0.983	0.957
		25	112	0.998	0.994	0.987	0.972	0.969
		50	106	0.985	0.978	0.973	0.965	0.951
		75	101	0.959	0.943	0.946	0.955	0.934
	2.5	11	57	1.000	0.999	0.998	0.993	0.985
		25	53	0.982	0.982	0.979	0.965	0.960
		50	49	0.597	0.678	0.726	0.735	0.749
		75	43	0.009	0.036	0.082	0.152	0.242
	5.0	11	37	0.999	0.999	0.999	0.995	0.989
		25	34	0.796	0.843	0.856	0.860	0.877
		50	29	0.000	0.000	0.000	0.000	0.001
		75	24	0.000	0.000	0.000	0.000	0.000
	7.5	11	30	0.995	0.998	0.994	0.991	0.991
		25	28	0.353	0.349	0.485	0.571	0.627
		50	22	0.000	0.000	0.000	0.000	0.000
		75	17	0.000	0.000	0.000	0.000	0.000
	10.0	11	27	0.992	0.996	0.997	0.992	0.987
		25	24	0.000	0.015	0.034	0.085	0.202
		50	19	0.000	0.000	0.000	0.000	0.000
		75	14	0.000	0.000	0.000	0.000	0.000

Acknowledgements

This research received financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grants 153526/2014-9 and 3304126/2015-2) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant 2013/21728-2), Brazil.

References

Adcock, C. J. (1987). A Bayesian approach to calculating sample sizes for multinomial sampling. *Journal of the Royal Statistical Society: Series D (The Statistician)* **36**, 155–159.

Adcock, C. J. (1988). A Bayesian approach to calculating sample sizes. *Journal of the Royal Statistical Society: Series D (The Statistician)* **37**, 433–439.

Adcock, C. J. (1997). Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)* **46**, 261–283.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152–1174.

Bailey, S. A. (2015). An overview of thirty years of research on ballast water as a vector for aquatic invasive species to freshwater and marine environments. *Aquatic Ecosystem Health & Management* **18**, 261–268.

Carney, K. J., Basurko, O. C., Pazouki, K., Marsham, S., Delany, J. E., Desai, D. V., Anil, A. C. & Mesbahi, E. (2013). Difficulties in obtaining representative samples for compliance with the Ballast Water Management Convention. *Marine Pollution Bulletin* **68**, 99–105.

Cohen, A. N., Dobbs, F. C. & Chapman, P. M. (2017). Revisiting the basis for us ballast water regulations. *Marine Pollution Bulletin* **118**, 348–353.

Costa, E. G. (2017). *Tamanho amostral para estimar a concentração de organismos em água de lastro: uma abordagem bayesiana*. Ph.D. thesis. Departamento de Estatística, Universidade de São Paulo, São Paulo. In Portuguese.

Costa, E. G., Lopes, R. M. & Singer, J. M. (2015). Implications of heterogeneous distributions of organisms on ballast water sampling. *Marine Pollution Bulletin* **91**, 280–287.

Costa, E. G., Lopes, R. M. & Singer, J. M. (2016). Sample size for estimating the mean concentration of organisms in ballast water. *Journal of Environmental Management* **180**, 433–438.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **170**, 95–113.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

First, M. R., Robbins-Wamsley, S. H., Riley, S. C., Moser, C. S., Smith, G. E., Tamburri, M. N. & Drake, L. A. (2013). Stratification of living organisms in ballast tanks: how do organism concentrations vary as ballast water is discharged? *Environmental Science & Technology* **47**, 4442–4448.

Gollasch, S. & David, M. (2017). Recommendations for representative ballast water sampling. *Journal of Sea Research* **123**, 1–15.

Inoue, L. Y., Berry, D. A. & Parmigiani, G. (2005). Relationship between bayesian and frequentist sample size determination. *The American Statistician* **59**, 79–87.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994a). *Continuous univariate distributions*, 2nd ed., vol. 1. New York: John Wiley & Sons.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994b). *Continuous univariate distributions*, 2nd ed., vol. 2. New York: John Wiley & Sons.

Joseph, L. & Bélisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society: Series D (The Statistician)* **46**, 209–226.

Joseph, L., Berger, R. D. & Bélisle, P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine* **16**, 769–781.

Joseph, L. & Wolfson, D. B. (1997). Interval-based versus decision theoretic criteria for the choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)* **46**, 145–149.

Joseph, L., Wolfson, D. B. & Berger, R. D. (1995). Sample size calculations for binomial proportions via highest posterior density intervals. *Journal of the Royal Statistical Society: Series D (The Statistician)* **44**, 143–154.

Ludwig, J. A. & Reynolds, J. F. (1988). *Statistical ecology: a primer on methods and computing*, 1st ed., vol. 1. John Wiley & Sons.

Marbuah, G., Gren, I.-M. & McKie, B. (2014). Economics of harmful invasive species: a review. *Diversity* **6**, 500–523.

Matuszewski, D. J., Cesar, R. M., Strickler, J. R., Baldasso, L. F. & Lopes, R. M. (2015). Visual rhythm for particle analysis in sample-in-flow systems: application for continuous plankton monitoring. *Limnology and Oceanography: Methods* **13**, 687–696.

M'Lan, C. E., Joseph, L. & Wolfson, D. B. (2006). Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association* **101**, 760–772.

M'Lan, C. E., Joseph, L. & Wolfson, D. B. (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis* **3**, 269–296.

Murphy, K. R., Ritz, D. & Hewitt, C. L. (2002). Heterogeneous zooplankton distribution in a ship's ballast tanks. *Journal of Plankton Research* **24**, 729–734.

Nassar, M. M., Khamis, S. M. & Radwan, S. S. (2010). Geometric sample size determination in Bayesian analysis. *Journal of Applied Statistics* **37**, 567–575.

Nassar, M. M., Khamis, S. M. & Radwan, S. S. (2011). On Bayesian sample size determination. *Journal of Applied Statistics* **38**, 1045–1054.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rahme, E., Joseph, L. & Gyorkos, T. W. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49**, 119–128.

Stamey, J. D., Young, D. M. & Bratcher, T. L. (2006). Bayesian sample-size determination for one and two Poisson rate parameters with applications to quality control. *Journal of Applied Statistics* **33**, 583–594.

Wang, F. & Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* **17**, 193–208.

White, G. C. & Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology* **77**, 2549–2557.

Supplementary Material

1 Algorithms to obtain n

A possible algorithm to obtain the sample size n is described as follows for each criterion and model. For simplicity of notation, we drop the argument $\boldsymbol{x_n}$ in the limits of the HPD interval, $a(\boldsymbol{x}_n)$ and $b(\boldsymbol{x}_n)$, for the posterior distribution throughout the text

1.1 Poisson/gamma model

1.1.1 Posterior distribution and properties

Under the Poisson/gamma model each X_i follows marginally a negative binomial distribution with mean $w\lambda_0$ and parameter θ_0, *i.e.*, $X_i \sim NB(w\lambda_0, \theta_0)$. Furthermore, $S_n = \sum_{i=1}^{n} X_i \sim NB(nw\lambda_0, n\theta_0)$.

The corresponding likelihood function is

$$L(\lambda; \boldsymbol{x}_n) = \prod_{i=1}^{n} \frac{e^{-w\lambda}(w\lambda)^{x_i}}{x_i!} = \frac{e^{-nw\lambda}(w\lambda)^{s_n}}{\prod_{i=1}^{n} x_i!},$$

where $s_n = \sum_{i=1}^{n} x_i$ and $\boldsymbol{x}_n = (x_1, \ldots, x_n)$. If we consider a prior gamma

distribution for λ, the posterior distribution is

$$\begin{aligned} f(\lambda|\boldsymbol{x}_n) &\propto \lambda^{s_n}\exp(-nw\lambda)\times\lambda^{\theta_0-1}\exp\left(-\frac{\theta_0}{\lambda_0}\lambda\right) \\ &= \lambda^{\theta_0+s-1}\exp\left[-\left(nw+\frac{\theta_0}{\lambda_0}\right)\lambda\right], \end{aligned} \tag{1}$$

which is a gamma distribution with parameters θ_0+s_n and $nw+\theta_0/\lambda_0$. The corresponding mean and variance are, respectively,

$$\mathbb{E}\left[\lambda\middle|\boldsymbol{x}_n\right]=\frac{\theta_0+s_n}{\theta_0/\lambda_0+nw} \quad \text{and} \quad \text{Var}\left[\lambda\middle|\boldsymbol{x}_n\right]=\frac{\theta_0+s_n}{(\theta_0/\lambda_0+nw)^2}. \tag{2}$$

To obtain the HPD interval of length ℓ for a gamma distribution with parameters κ and ψ, we denote as $G(\kappa,\psi)$, it is necessary to solve the equation $g(a)=g(a+\ell)$ in a, where $g(\cdot)$ is the corresponding density function (Stamey et al., 2006). The solution is

$$a=\frac{\ell}{\exp(\psi\ell/(\kappa-1))-1}. \tag{3}$$

For the posterior distribution of the Poisson/gamma model we consider $\kappa=\theta_0+s_n$ and $\psi=nw+\theta_0/\lambda_0$.

If we want to obtain the HPD interval with probability $1-\rho$ we need to solve a system of equations $g(a)=g(b)$ and $\mathbb{P}\left[a\leq\lambda\leq b|s_n\right]=1-\rho$, where a and b are the lower and upper limits of the HPD interval, respectively. To solve this system we need to use numerical methods (*e.g.*, Newton-Raphson).

1.1.2 Average coverage criterion algorithm

Step 1. Set values for ℓ, λ_0, θ_0, w, ρ and take $n = 1$;

Step 2. Draw a sample of size M (*e.g.*, $M = 1000$) of s_n from a negative binomial distribution with mean $nw\lambda_0$ and shape parameter $n\theta_0$; the size of replicates M must be as large as possible in order to enhance the estimate of the integral (or expected value) associated with the criterion;

Step 3. Compute the lower limit a of the HPD interval using (3) and the posterior probability $\mathbb{P}[a \leq \lambda \leq a + \ell | s_n]$ using (1) for each s_n that was drawn: for each value drawn in Step 2 compute $\theta_0 + s_n$ and $nw + \theta_0/\lambda_0$ corresponding to the posterior gamma distribution of λ, then compute the lower limit a of the HPD interval using (3). Then, compute the posterior probabilities using the posterior gamma distribution with parameters $\theta_0 + s_n$ and $nw + \theta_0/\lambda_0$ (1);

Step 4. Compute the average of the M posterior probabilities;

Step 5. If this average is $\geq 1 - \rho$, stop. The value n obtained in this step is the required value. Otherwise, set $n = n + 1$ and return to Step 2.

1.1.3 Average length criterion algorithm

Step 1. Set values for $\ell_{\max}$, λ_0, θ_0, w, ρ and take $n = 1$;

Step 2. Draw a sample of size M (*e.g.*, $M = 1000$) of s_n from a negative binomial distribution with mean $nw\lambda_0$ and shape parameter $n\theta_0$;

Step 3. Compute the length of the HPD interval of probability $1-\rho$ for each s_n that was drawn: for each value drawn in Step 2 compute $\theta_0 + s_n$ and $nw+\theta_0/\lambda_0$ corresponding to the posterior gamma distribution of λ, and obtain the lower and upper limits of the HPD interval of probability $1-\rho$. Then, compute the difference between the upper and the lower limits for each value drawn in order to obtain the interval lengths;

Step 4. Compute the average of the M lengths of the HPD intervals;

Step 5. If this average is $\leq \ell_{\max}$, stop. The value n obtained in this step is the required value. Otherwise, set $n = n + 1$ and return to Step 2.

1.2 Negative binomial/Pearson Type VI model

1.2.1 Posterior distribution and properties

Consider the negative binomial model. The corresponding likelihood function is

$$\begin{aligned} L(\lambda; \boldsymbol{x}_n) &= \prod_{i=1}^{n} \frac{\Gamma(\phi + x_i)}{\Gamma(x_i+1)\Gamma(\phi)} \left(\frac{w\lambda}{w\lambda+\phi}\right)^{x_i} \left(\frac{\phi}{w\lambda+\phi}\right)^{\phi} \\ &= \left[\prod_{i=1}^{n} \frac{\Gamma(\phi + x_i)}{\Gamma(x_i+1)\Gamma(\phi)}\right] \left(\frac{w}{\phi}\lambda\right)^{s_n} \left(1+\frac{w}{\phi}\lambda\right)^{-s_n-n\phi}, \end{aligned}$$

where $s_n = \sum_{i=1}^{n} x_i$ and $\boldsymbol{x}_n = (x_1, \ldots, x_n)$. If we consider a Pearson Type VI prior distribution for λ, the posterior distribution is

$$\begin{aligned} f(\lambda|\boldsymbol{x}_n) &\propto \left(\frac{w}{\phi}\lambda\right)^{s_n} \left(1+\frac{w}{\phi}\lambda\right)^{-s_n-n\phi} \times \left(\frac{w}{\phi}\lambda\right)^{\theta_0-1} \left(1+\frac{w}{\phi}\lambda\right)^{-\theta_0-(\theta_0/\lambda_0+1)} \\ &= \left(\frac{w}{\phi}\lambda\right)^{\theta_0+s_n-1} \left(1+\frac{w}{\phi}\lambda\right)^{-(\theta_0+s_n)-(\theta_0/\lambda_0+n\phi+1)}, \end{aligned}$$

which corresponds to a Pearson Type VI distribution with location and scale parameters equal to 0 and ϕ/w, respectively, and shape parameters $\theta_0 + s_n$ and $\theta_0/\lambda_0 + n\phi + 1$. The corresponding mean and variance is

$$\mathbb{E}\left[\lambda\middle|\boldsymbol{x}_n\right] = \frac{\phi}{w}\frac{\theta_0 + s_n}{\theta_0/\lambda_0 + n\phi} \quad \text{and} \quad \text{Var}\left[\lambda\middle|\boldsymbol{x}_n\right] = \left(\frac{\phi}{w}\right)^2 \frac{\theta_0 + s_n}{(\theta_0/\lambda_0 + n\phi)^2}\left(\frac{\mathbb{E}\left[\lambda\middle|\boldsymbol{x}_n\right] + 1}{1-q}\right), \tag{4}$$

where $q = (\theta_0/\lambda_0 + n\phi)^{-1}$. To obtain the HPD interval of length ℓ for the Pearson Type VI distribution, we denote as $PVI(0, \phi/w, \kappa, \psi)$, it is necessary to solve the equation $g(a) = g(a + \ell)$ in a, where $g(\cdot)$ is the corresponding density function. The solution is obtained from the equation

$$(\kappa - 1)\log\left[1 + \frac{\ell}{a}\right] - (\kappa + \psi)\log\left[1 + \frac{w\ell}{\phi + wa}\right] = 0, \tag{5}$$

which may be solved by numerical methods (*e.g.*, Newton-Raphson). For the posterior distribution obtained from the negative binomial/Pearson Type VI model we have $\kappa = \theta_0 + s_n$ and $\psi = \theta_0/\lambda_0 + n\phi + 1$.

If we want to obtain the HPD interval with probability $1 - \rho$ we need to solve a system of equations $g(a) = g(b)$ and $\mathbb{P}\left[a \leq \lambda \leq b|s_n\right] = 1 - \rho$, where a and b are the lower and upper limits of the HPD interval, respectively. To solve this system we also need to use numerical methods (*e.g.*, Newton-Raphson).

1.2.2 Average coverage criterion algorithm

Step 1. Set values for ℓ, ϕ, λ_0, θ_0, w, ρ and take $n = 1$;

Step 2. Draw a sample of size M (*e.g.*, $M = 1000$) of s_n; to draw s_n, first

draw a sample of size n of λ from the prior distribution $PVI(0, \phi/w, \theta_0, \theta_0/\lambda_0+1)$, and given these values draw a sample of size n of X_i from the negative binomial distribution with mean $w\lambda$ and shape parameter ϕ, then add the X_i's that were drawn;

Step 3. Compute the lower limit a of the HPD interval using (5) and the posterior probability $\mathbb{P}[a \leq \lambda \leq a + \ell | s_n]$ using (4) for each s_n that was drawn: for each value drawn in Step 2, compute $\theta_0 + s_n$ and $\theta_0/\lambda_0 + n\phi + 1$ corresponding to the Pearson Type VI posterior distribution of λ, then compute the lower limit a of the HPD interval using (5). Then, compute the posterior probabilities using the Pearson Type VI distribution with parameters $\theta_0 + s_n$ and $\theta_0/\lambda_0 + n\phi + 1$ (1);

Step 4. Compute the average of the M posterior probabilities;

Step 5. If this average is $\geq 1 - \rho$, stop. The value n obtained in this step is the required value. Otherwise, set $n = n + 1$ and return to Step 2.

1.2.3 Average length criterion algorithm

Step 1. Set values for $\ell_{\max}$, ϕ, λ_0, θ_0, w, ρ and take $n = 1$;

Step 2. Draw a sample of size M (*e.g.*, $M = 1000$) of s_n; to draw s_n, first draw a sample of size n of λ from the prior distribution $PVI(0, \phi/w, \theta_0, \theta_0/\lambda_0+1)$, and given this values, draw a sample of size n of X_i from the negative binomial distribution with mean $w\lambda$ and shape parameter ϕ, then add the X_i's that were drawn;

Step 3. Compute the length of the HPD interval of probability $1-\rho$ for each s_n that was drawn: for each value drawn in Step 2, compute $\theta_0 + s_n$ and $\theta_0/\lambda_0 + n\phi + 1$ corresponding to the Pearson Type VI posterior distribution of λ, then obtain the lower and upper limits of the HPD interval of probability $1 - \rho$. Then, compute the difference between the upper and lower limits for each value drawn in order to obtain the interval lengths;

Step 4. Compute the average of the M HPD interval lengths;

Step 5. If this average is $\leq \ell_{\max}$, stop. The value n obtained in this step is the required value. Otherwise, set $n = n + 1$ and return to Step 2.

2 Proof of Theorem 1

Before presenting the proof of Theorem 1 we need some results.

Lemma 1. *Let* $\{X_n\}_{n\geq 1}$ *be a sequence of random variables. If* $\sup_n \mathbb{E}\left[|X_n|^{1+\epsilon}\right] < \infty$ *for some* $\epsilon > 0$*, then* X_n *is uniformly integrable.*

Proof. See Billingsley (1995, p. 338). □

Theorem 2. *If* $X_n \to X$ *in distribution as* $n \to \infty$ *and* X_n *is uniformly integrable, then* X *is integrable and* $\mathbb{E}\left[X_n\right] \to \mathbb{E}\left[X\right]$*, as* $n \to \infty$*.*

Proof. See Billingsley (1995, Theorem 25.12). □

Theorem 3. *Suppose that* $X_n \to X$ *in distribution as* $n \to \infty$ *and that* $h_n(x)$ *and* $h(x)$ *are Borel functions. Let* $E \subseteq \mathbb{R}$ *be a set in which, for some sequence* $x_n \to x$*, the statement* $h_n(x) \to h(x)$ *as* $n \to \infty$ *does not hold. Suppose that* $\mathbb{P}\left[X(\omega) \in E\right] = 0$*. Then,* $h_n(X_n) \to h(X)$ *in distribution as* $n \to \infty$*.*

Proof. See Billingsley (1995, p. 340). □

Consider the Poisson/gamma model (1)-(2) and the average length criterion, then we have that $\ell'(\boldsymbol{x}_n, n)$ may be approximated by $2z_{\rho/2}\sqrt{\text{Var}\left[\lambda\middle|\boldsymbol{x}_n\right]}$, where $\text{Var}\left[\lambda\middle|\boldsymbol{x}_n\right] = (\theta_0 + s_n)/(nw + \theta_0/\lambda_0)^2$ and $z_{\rho/2}$ is the quantile of probability $1-\rho/2$ of the standard normal distribution, since the posterior variance may be implicitly related with the length of the corresponding HPD interval (Joseph et al., 1995).

Proposition 1. *Consider the Poisson/gamma model (1)-(2) and define* $S_n = \sum_{i=1}^n X_i$. *Then,*

$$\lim_{n\to\infty} \left(nw + \frac{\theta_0}{\lambda_0}\right)^{1/2} \frac{\mathbb{E}\left[\ell'(\boldsymbol{X}_n, n)\right]}{2z_{\rho/2}} = \frac{\Gamma(\theta_0 + 1/2)}{\Gamma(\theta_0)} \left(\frac{\lambda_0}{\theta_0}\right)^{1/2}. \tag{6}$$

Proof. Define $Y_n = S_n/nw$ and let $\mathcal{F}$ be the set of points in which Y_n has positive probability. Using Theorem B.1 of M'Lan et al. (2006), we have $S_n/n - w\lambda \to 0$ in probability as $n \to \infty$. Then, $Y_n - \lambda \to 0$ in distribution as $n \to \infty$ (see Sen et al., 2009, Theorem 6.2.7).

Let $h_n(y) = \left[\frac{\theta_0 + nwy}{nw + \theta_0/\lambda_0}\right]^{1/2}$ and $h(y) = y^{1/2} = \lim_{n\to\infty} h_n(y)$. Consider the Lemma 1 with $\epsilon = 1$; then,

$$\begin{aligned}
\mathbb{E}\left[|Y_n|^{1+\epsilon}\right] &= \frac{\mathbb{E}\left[|S_n|^{1+\epsilon}\right]}{(nw)^{1+\epsilon}} = \frac{\mathbb{E}\left[|S_n|^2\right]}{(nw)^2} = \frac{1}{(nw)^2}[\mathrm{Var}\left[S_n\right] + (\mathbb{E}\left[S_n\right])^2] \\
&= \frac{1}{(nw)^2}\left[nw\lambda_0 + \frac{n(w\lambda_0)^2}{\theta_0} + (nw\lambda_0)^2\right] = \frac{\lambda_0}{nw} + \lambda_0^2\left(1 + \frac{1}{n\theta_0}\right).
\end{aligned}$$

Thus,

$$\sup_n \mathbb{E}\left[|Y_n|^{1+\epsilon}\right] = \sup_n \left[\frac{\lambda_0}{nw} + \lambda_0^2\left(1 + \frac{1}{n\theta_0}\right)\right] = \frac{\lambda_0}{w} + \lambda_0^2\left(1 + \frac{1}{\theta_0}\right) < \infty,$$

and by Lemma 1, Y_n is uniformly integrable. Using Theorem 2, we have $\mathbb{E}\left[Y_n\right] \to \mathbb{E}\left[\lambda\right]$ as $n \to \infty$ and by Theorem 3, $h_n(Y_n) \to h(\lambda)$ in distribution

as $n \to \infty$. Then, we have

$$\left(nw + \frac{\theta_0}{\lambda_0}\right)^{1/2} \frac{\mathbb{E}\left[\ell'(\boldsymbol{X}_n, n)\right]}{2z_{\rho/2}} \cong \sum_{s_n=0}^{\infty} \left[\frac{\theta_0 + s_n}{nw + \theta_0/\lambda_0}\right]^{1/2} f_{S_n}(s_n) = \sum_{y_n \in \mathcal{F}} h_n(y_n) f_{Y_n}(y_n)$$

$$= \mathbb{E}\left[h(Y_n)\right] \to \mathbb{E}\left[h(\lambda)\right]$$

$$= \int_0^{\infty} h(\lambda) \frac{(\theta_0/\lambda_0)^{\theta_0}}{\Gamma(\theta_0)} \lambda^{\theta_0 - 1} \exp(-\theta_0 \lambda/\lambda_0) d\lambda$$

$$= \int_0^{\infty} \frac{(\theta_0/\lambda_0)^{\theta_0}}{\Gamma(\theta_0)} \lambda^{\theta_0 + 1/2 - 1} \exp(-\theta_0 \lambda/\lambda_0) d\lambda$$

$$= \left(\frac{\lambda_0}{\theta_0}\right)^{1/2} \frac{\Gamma(\theta_0 + 1/2)}{\Gamma(\theta_0)}. \qquad \square$$

Finally, we consider the proof of Theorem 1. According to Proposition 1, the average length criterion may be approximated by (6). Thus, it is enough to obtain the smallest n which satisfies

$$\frac{2z_{\rho/2}}{(nw + \theta_0/\lambda_0)^{1/2}} \frac{\Gamma(\theta_0 + 1/2)}{\Gamma(\theta_0)} \left(\frac{\lambda_0}{\theta_0}\right)^{1/2} \leq \ell_{\max},$$

and solving the inequality in n we obtain the result of the Theorem 1.

3 Tables

In some cells of Table S6 the symbol "-" means that in this case any simulated HPD interval contains the fixed value of λ.

Table S1: ACC based Bayesian coverage probability of the HPD interval estimated through simulation for some scenarios under the model (3)-(4) using the sample sizes in Table 2 with $w = 1$.

Interval length (ℓ)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	426	1.00	1.00	1.00	0.97	0.85
		25	422	1.00	1.00	0.99	0.96	0.84
		50	417	1.00	0.99	0.97	0.93	0.84
		75	414	0.99	0.99	0.97	0.94	0.85
	2.5	11	194	1.00	1.00	0.99	0.95	0.82
		25	191	1.00	0.99	0.97	0.94	0.86
		50	188	0.99	0.97	0.98	0.93	0.88
		75	185	0.99	0.98	0.96	0.94	0.91
	5.0	11	115	1.00	0.99	0.98	0.94	0.86
		25	113	1.00	0.99	0.97	0.93	0.88
		50	111	0.98	0.97	0.97	0.94	0.89
		75	108	0.97	0.98	0.96	0.95	0.92
	7.5	11	90	1.00	0.99	0.97	0.93	0.86
		25	88	0.99	0.98	0.97	0.94	0.90
		50	85	0.98	0.97	0.95	0.94	0.90
		75	82	0.97	0.96	0.96	0.97	0.92
	10.0	11	77	0.99	0.98	0.97	0.92	0.89
		25	75	0.99	0.97	0.97	0.95	0.90
		50	72	0.98	0.98	0.96	0.94	0.92
		75	70	0.98	0.97	0.96	0.95	0.92
4	1.0	11	110	1.00	1.00	1.00	0.99	0.86
		25	107	1.00	0.99	0.99	0.97	0.86
		50	102	0.99	0.99	0.98	0.95	0.88
		75	99	0.99	0.99	0.98	0.96	0.88
	2.5	11	49	1.00	1.00	0.99	0.97	0.88
		25	46	0.99	0.98	0.97	0.96	0.88
		50	44	0.98	0.98	0.97	0.95	0.90
		75	41	0.98	0.98	0.98	0.97	0.92
	5.0	11	29	0.99	0.98	0.99	0.95	0.87
		25	27	0.98	0.99	0.98	0.96	0.89
		50	24	0.97	0.97	0.98	0.97	0.93
		75	22	0.97	0.97	0.99	0.97	0.94
	7.5	11	22	0.99	0.98	0.96	0.94	0.89
		25	20	0.98	0.99	0.97	0.96	0.90
		50	18	0.98	0.98	0.98	0.97	0.94
		75	15	0.96	0.99	0.98	0.98	0.95
	10.0	11	19	0.99	0.98	0.98	0.95	0.90
		25	17	0.98	0.98	0.98	0.96	0.93
		50	15	0.98	0.99	0.98	0.98	0.96
		75	12	0.97	0.99	0.99	0.98	0.96

Table S2: ALC based length of the HPD interval estimated through simulation for some scenarios under the model (3)-(4) using the sample sizes in Table 4 with $w = 0.5$.

Maximum interval length ($\ell_{\max}$)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	456	0.91	1.20	1.54	2.00	2.85
		25	456	1.21	1.48	1.76	2.12	2.71
		50	452	1.42	1.65	1.87	2.14	2.55
		75	445	1.52	1.73	1.92	2.14	2.46
	2.5	11	228	1.22	1.50	1.77	2.12	2.70
		25	226	1.47	1.68	1.89	2.14	2.50
		50	221	1.62	1.79	1.95	2.12	2.36
		75	216	1.69	1.84	1.96	2.10	2.29
	5.0	11	151	1.40	1.65	1.87	2.14	2.55
		25	150	1.60	1.77	1.93	2.11	2.37
		50	144	1.72	1.85	1.97	2.09	2.26
		75	138	1.79	1.89	1.99	2.08	2.22
	7.5	11	126	1.48	1.70	1.91	2.14	2.48
		25	123	1.66	1.82	1.96	2.11	2.32
		50	118	1.77	1.88	1.98	2.08	2.23
		75	113	1.81	1.91	1.99	2.07	2.18
	10.0	11	114	1.52	1.72	1.91	2.12	2.42
		25	110	1.69	1.84	1.97	2.11	2.30
		50	106	1.79	1.89	1.97	2.07	2.19
		75	101	1.83	1.91	1.98	2.05	2.15
4	1.0	11	113	1.84	2.42	3.06	3.98	5.69
		25	109	2.47	3.00	3.54	4.27	5.43
		50	105	2.89	3.33	3.76	4.28	5.04
		75	99	3.14	3.51	3.85	4.27	4.86
	2.5	11	55	2.49	3.03	3.58	4.24	5.37
		25	52	3.00	3.42	3.81	4.29	4.98
		50	48	3.31	3.62	3.89	4.21	4.63
		75	43	3.48	3.73	3.91	4.15	4.47
	5.0	11	37	2.81	3.28	3.71	4.23	5.01
		25	34	3.24	3.57	3.85	4.19	4.66
		50	29	3.52	3.73	3.91	4.13	4.40
		75	23	3.71	3.86	3.98	4.13	4.32
	7.5	11	30	2.99	3.41	3.80	4.24	4.89
		25	27	3.37	3.66	3.92	4.19	4.59
		50	22	3.63	3.80	3.95	4.13	4.33
		75	17	3.76	3.87	3.97	4.09	4.22
	10.0	11	27	3.07	3.45	3.81	4.21	4.79
		25	24	3.43	3.69	3.92	4.16	4.50
		50	19	3.66	3.82	3.94	4.10	4.28
		75	14	3.78	3.88	3.96	4.05	4.16

Table S3: ALC based length of the HPD interval estimated through simulation for some scenarios under the model (3)-(4) using the sample sizes in Table 4 with $w = 1$.

Maximum interval length ($\ell_{\max}$)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	414	0.66	0.92	1.25	1.74	2.79
		25	418	0.97	1.26	1.58	2.02	2.84
		50	416	1.22	1.49	1.77	2.12	2.71
		75	410	1.35	1.60	1.85	2.15	2.62
	2.5	11	191	0.99	1.28	1.60	2.04	2.81
		25	190	1.28	1.54	1.80	2.13	2.66
		50	186	1.48	1.70	1.91	2.14	2.50
		75	185	1.57	1.75	1.93	2.13	2.40
	5.0	11	115	1.22	1.49	1.77	2.12	2.68
		25	113	1.46	1.69	1.90	2.14	2.50
		50	110	1.63	1.80	1.94	2.12	2.36
		75	108	1.69	1.84	1.96	2.10	2.28
	7.5	11	89	1.33	1.58	1.83	2.13	2.60
		25	88	1.54	1.74	1.92	2.12	2.41
		50	85	1.68	1.83	1.96	2.10	2.29
		75	82	1.75	1.87	1.98	2.08	2.24
	10.0	11	76	1.40	1.64	1.87	2.14	2.54
		25	75	1.59	1.77	1.94	2.11	2.37
		50	72	1.73	1.85	1.97	2.09	2.26
		75	70	1.78	1.88	1.97	2.07	2.19
4	1.0	11	103	1.33	1.86	2.49	3.46	5.54
		25	104	1.94	2.52	3.14	4.00	5.59
		50	100	2.48	3.00	3.55	4.23	5.34
		75	97	2.77	3.25	3.70	4.28	5.15
	2.5	11	47	2.00	2.57	3.20	4.08	5.59
		25	46	2.59	3.09	3.60	4.22	5.21
		50	43	3.03	3.42	3.81	4.22	4.87
		75	41	3.23	3.55	3.84	4.18	4.68
	5.0	11	28	2.46	3.00	3.54	4.20	5.29
		25	26	3.01	3.42	3.82	4.27	4.94
		50	24	3.34	3.62	3.88	4.18	4.59
		75	22	3.46	3.68	3.90	4.10	4.40
	7.5	11	21	2.72	3.21	3.69	4.27	5.19
		25	20	3.16	3.51	3.86	4.20	4.74
		50	18	3.42	3.66	3.88	4.11	4.41
		75	15	3.63	3.81	3.93	4.11	4.31
	10.0	11	19	2.77	3.22	3.66	4.16	4.93
		25	17	3.26	3.57	3.85	4.19	4.63
		50	15	3.48	3.69	3.87	4.03	4.31
		75	12	3.67	3.80	3.93	4.06	4.23

Table S4: ACC based Bayesian coverage probability of the HPD interval estimated through simulation of data using $\phi = 0.5$ and applying the methodology with the scenarios specified bellow under the model (3)-(4) using the sample sizes in Table 2 with $w = 1$.

Interval length (ℓ)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	426	1.000	1.000	1.000	1.000	0.947
		25	422	1.000	1.000	0.997	0.986	0.947
		50	417	1.000	0.997	0.992	0.987	0.959
		75	414	0.998	0.992	0.991	0.979	0.967
	2.5	11	194	1.000	1.000	1.000	1.000	0.998
		25	191	1.000	1.000	0.999	0.992	0.994
		50	188	1.000	1.000	0.991	0.995	0.989
		75	185	0.989	0.996	0.986	0.986	0.980
	5.0	11	115	1.000	1.000	1.000	1.000	0.997
		25	113	1.000	1.000	1.000	0.998	0.998
		50	111	0.997	0.994	0.989	0.989	0.991
		75	108	0.898	0.904	0.905	0.909	0.898
	7.5	11	90	1.000	1.000	1.000	1.000	1.000
		25	88	1.000	1.000	1.000	0.998	0.998
		50	85	0.967	0.968	0.968	0.955	0.954
		75	82	0.319	0.360	0.452	0.483	0.493
	10.0	11	77	1.000	1.000	1.000	1.000	1.000
		25	75	1.000	0.999	1.000	1.000	0.999
		50	72	0.810	0.807	0.819	0.865	0.865
		75	70	0.004	0.012	0.024	0.035	0.077
4	1.0	11	110	1.000	0.999	0.999	0.994	0.972
		25	107	1.000	0.996	0.997	0.991	0.962
		50	102	0.996	0.988	0.987	0.972	0.958
		75	99	0.990	0.980	0.984	0.961	0.955
	2.5	11	49	1.000	1.000	0.999	0.998	0.977
		25	46	0.996	0.996	0.992	0.988	0.971
		50	44	0.957	0.939	0.938	0.948	0.928
		75	41	0.671	0.655	0.768	0.771	0.826
	5.0	11	29	0.999	1.000	0.996	0.995	0.992
		25	27	0.976	0.976	0.973	0.967	0.962
		50	24	0.115	0.207	0.335	0.406	0.454
		75	22	0.000	0.000	0.000	0.000	0.000
	7.5	11	22	1.000	1.000	0.997	0.997	0.993
		25	20	0.831	0.858	0.877	0.883	0.878
		50	18	0.000	0.000	0.000	0.000	0.000
		75	15	0.000	0.000	0.000	0.000	0.000
	10.0	11	19	1.000	0.998	0.996	0.998	0.994
		25	17	0.442	0.563	0.671	0.733	0.716
		50	15	0.000	0.000	0.000	0.000	0.000
		75	12	0.000	0.000	0.000	0.000	0.000

Table S5: ALC based length of the HPD interval estimated through simulation of data using $\phi = 0.5$ and applying the methodology with the scenarios specified below under the model (3)-(4) using the sample sizes in Table 4 with $w = 0.5$.

Maximum interval length ($\ell_{\max}$)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	456	0.528	0.677	0.847	1.077	1.515
		25	456	0.683	0.824	0.966	1.150	1.447
		50	452	0.799	0.913	1.026	1.163	1.373
		75	445	0.856	0.959	1.054	1.167	1.332
	2.5	11	228	0.427	0.497	0.566	0.648	0.777
		25	226	0.499	0.553	0.603	0.657	0.742
		50	221	0.551	0.593	0.628	0.670	0.726
		75	216	0.575	0.612	0.643	0.678	0.724
	5.0	11	151	0.382	0.425	0.466	0.509	0.571
		25	150	0.432	0.466	0.494	0.524	0.562
		50	144	0.476	0.495	0.519	0.535	0.565
		75	138	-	-	-	-	-
	7.5	11	126	0.358	0.394	0.423	0.454	0.501
		25	123	0.407	0.430	0.451	0.474	0.505
		50	118	-	-	-	-	0.512
		75	113	-	-	-	-	-
	10.0	11	114	0.341	0.369	0.395	0.424	0.459
		25	110	0.389	0.408	0.428	0.443	0.466
		50	106	-	-	-	-	-
		75	101	-	-	-	-	-
4	1.0	11	113	1.082	1.381	1.704	2.177	3.024
		25	109	1.435	1.697	1.988	2.339	2.911
		50	105	1.702	1.922	2.148	2.409	2.790
		75	99	1.860	2.049	2.252	2.459	2.764
	2.5	11	55	0.921	1.048	1.183	1.341	1.601
		25	52	1.126	1.230	1.320	1.435	1.592
		50	48	-	1.322	1.402	1.495	1.606
		75	43	-	-	-	-	1.636
	5.0	11	37	0.870	0.949	1.019	1.098	1.224
		25	34	-	-	1.139	1.189	1.275
		50	29	-	-	-	-	-
		75	23	-	-	-	-	-
	7.5	11	30	0.863	0.940	0.997	1.051	1.138
		25	27	-	-	-	-	-
		50	22	-	-	-	-	-
		75	17	-	-	-	-	-
	10.0	11	27	-	0.916	0.942	1.008	1.066
		25	24	-	-	-	-	-
		50	19	-	-	-	-	-
		75	14	-	-	-	-	-

Table S6: ALC based length of the HPD interval estimated through simulation of data using $\phi = 0.5$ and applying the methodology with the scenarios specified below under the model (3)-(4) using the sample sizes in Table 4 with $w = 1$.

Maximum interval length ($\ell_{\max}$)	ϕ	θ_0	n	Probability quantile used to fix λ				
				1/6	2/6	3/6	4/6	5/6
2	1.0	11	414	0.658	0.921	1.247	1.744	2.794
		25	418	0.966	1.253	1.577	2.026	2.832
		50	416	1.213	1.485	1.765	2.124	2.716
		75	410	1.351	1.608	1.847	2.141	2.624
	2.5	11	191	0.990	1.289	1.605	2.038	2.831
		25	190	1.276	1.538	1.808	2.129	2.659
		50	186	1.483	1.702	1.901	2.140	2.515
		75	185	1.572	1.765	1.939	2.124	2.401
	5.0	11	115	1.220	1.489	1.760	2.111	2.691
		25	113	1.467	1.680	1.889	2.131	2.496
		50	110	1.635	1.789	1.968	2.126	2.356
		75	108	1.700	1.838	1.954	2.098	2.288
	7.5	11	89	1.327	1.579	1.843	2.132	2.606
		25	88	1.547	1.746	1.906	2.111	2.418
		50	85	1.689	1.827	1.966	2.102	2.278
		75	82	1.762	1.870	1.987	2.079	2.226
	10.0	11	76	1.397	1.634	1.877	2.138	2.540
		25	75	1.591	1.776	1.928	2.114	2.368
		50	72	1.728	1.852	1.971	2.077	2.264
		75	70	1.787	1.868	1.956	2.070	2.197
4	1.0	11	103	1.332	1.854	2.490	3.502	5.476
		25	104	1.950	2.493	3.094	4.009	5.619
		50	100	2.474	3.014	3.566	4.205	5.363
		75	97	2.757	3.227	3.706	4.252	5.214
	2.5	11	47	1.996	2.598	3.185	4.082	5.622
		25	46	2.586	3.071	3.592	4.224	5.202
		50	43	3.026	3.419	3.807	4.194	4.885
		75	41	3.233	3.572	3.836	4.194	4.659
	5.0	11	28	2.475	2.977	3.532	4.171	5.381
		25	26	3.017	3.389	3.813	4.241	4.926
		50	24	3.314	3.624	3.917	4.168	4.584
		75	22	3.487	3.693	3.914	4.105	4.386
	7.5	11	21	2.711	3.222	3.697	4.260	5.155
		25	20	3.143	3.502	3.777	4.246	4.773
		50	18	3.411	3.684	3.893	4.107	4.388
		75	15	3.599	3.804	3.916	4.086	4.322
	10.0	11	19	2.749	3.217	3.661	4.154	4.873
		25	17	3.221	3.526	3.867	4.188	4.678
		50	15	3.473	3.653	3.852	4.049	4.300
		75	12	3.666	3.833	3.974	4.083	4.186

References

Billingsley, P. (1995). *Probability and measure*. John Wiley & Sons, New York.

Joseph, L., Wolfson, D. B., and Berger, R. D. (1995). Sample size calculations for binomial proportions via highest posterior density intervals. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44:143–154.

M'Lan, C. E., Joseph, L., and Wolfson, D. B. (2006). Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association*, 101:760–772.

Sen, P. K., Singer, J. M., and de Lima, A. C. P. (2009). *From Finite Sample to Asymptotic Methods in Statistics*. Cambridge University Press, Cambridge.

Stamey, J. D., Young, D. M., and Bratcher, T. L. (2006). Bayesian sample-size determination for one and two Poisson rate parameters with applications to quality control. *Journal of Applied Statistics*, 33(6):583–594.